Chapter ML:VII (continued)

VII. Bayesian Learning

- Approaches to Probability
- Conditional Probability
- Bayes Classifier
- Exploitation of Data
- Frequentist versus Subjectivist
Exploitation of Data

Data Events

Data from a “predictor-response” setting:

\[ D = \{(x_1, y_1), \ldots, (x_n, y_n)\} \quad \text{(regression)} \]
\[ D = \{(x_1, c_1), \ldots, (x_n, c_n)\} \quad \text{(classification)} \]

- \( D \) is the result of \( n \) i.i.d. trials. I.e., \( n \) objects are sampled independently and from the same probability distribution. All objects are characterized by a “response” variable that is either quantitative (a number \( y \)) or categorical (a class label \( c \)), and by \( p \) “predictors” (a feature vector \( x \)).

- \( p(x_i, c_i), p(x_i, c_i) := P(X_i=x_i, C_i=c_i) \), is the probability of the joint event \( \{X_i=x_i, C_i=c_i\} \), i.e., (1) to get the vector \( x_i \), and, (2) that the respective object belongs to class \( c_i \). The \( p(x_i, y_i) \) are defined analogously.

- The \( Y_i \), \( C_i \), and \( X_i \) are i.i.d. (multivariate) random variables. Typically, the \( Y_i \) are of continuous type, the \( C_i \) of discrete type, and the variables of the random vector \( X_i \), \( X_i := (X_{1,i}, \ldots, X_{p,i})^T \), of continuous type.
Exploitation of Data

Data Events

Data from a “predictor-response” setting:

\[ D = \{(x_1, y_1), \ldots, (x_n, y_n)\} \quad \text{(regression)} \]

\[ D = \{(x_1, c_1), \ldots, (x_n, c_n)\} \quad \text{(classification)} \]

- \( D \) is the result of \( n \) i.i.d. trials. I.e., \( n \) objects are sampled independently and from the same probability distribution. All objects are characterized by a “response” variable that is either quantitative (a number \( y \)) or categorical (a class label \( c \)), and by \( p \) “predictors” (a feature vector \( x \)).

- \( p(x_i, c_i), p(x_i, c_i) := P(X_i=x_i, C_i=c_i) \), is the probability of the joint event \( \{X_i=x_i, C_i=c_i\} \), i.e., (1) to get the vector \( x_i \), and, (2) that the respective object belongs to class \( c_i \). The \( p(x_i, y_i) \) are defined analogously.

- The \( Y_i, C_i, \) and \( X_i \) are i.i.d. (multivariate) random variables. Typically, the \( Y_i \) are of continuous type, the \( C_i \) of discrete type, and the variables of the random vector \( X_i, X_i := (X_{1,i}, \ldots, X_{p,i})^T \), of continuous type.
Exploitation of Data

Data Events (continued)

Data from an “outcome-only” setting:

\[ D = \{ y_1, \ldots, y_n \} \quad \text{(quantitative)} \]

\[ D = \{ c_1, \ldots, c_n \} \quad \text{(categorical)} \]

- \( D \) is the result of \( n \) i.i.d. trials. I.e., \( n \) outcomes are sampled independently and from the same probability distribution. All outcomes are characterized by either a number \( y \) or a class label \( c \).

- \( p(y_i), p(y_i) := P(Y_i=y_i) \), is the probability of the event \( Y_i=y_i \). \( p(c_i), p(c_i) := P(C_i=c_i) \), is the probability of the event \( C_i=c_i \).

- The \( Y_i, \) and \( C_i \) are i.i.d. random variables. Typically, the \( Y_i \) are of continuous type and the \( C_i \) of discrete type.
Data Events (continued)

Data from an “outcome-only” setting:

\[ D = \{ y_1, \ldots, y_n \} \quad \text{(quantitative)} \]
\[ D = \{ c_1, \ldots, c_n \} \quad \text{(categorical)} \]

- \( D \) is the result of \( n \) i.i.d. trials. I.e., \( n \) outcomes are sampled independently and from the same probability distribution. All outcomes are characterized by either a number \( y \) or a class label \( c \).

- \( p(y_i), p(y_i) := P(Y_i = y_i) \), is the probability of the event \( Y_i = y_i \).
  
  \( p(c_i), p(c_i) := P(C_i = c_i) \), is the probability of the event \( C_i = c_i \).

- The \( Y_i \), and \( C_i \) are i.i.d. random variables. Typically, the \( Y_i \) are of continuous type and the \( C_i \) of discrete type.
Remarks:

- The following remarks on the predictor-response setting are detailed for a categorical response variable $c$; they apply to a quantitative response variable $y$ as well.

- By **experiment design**, the $n$ joint events, $\{X_1=x_1, C_1=c_1\}, \ldots, \{X_n=x_n, C_n=c_n\}$, generating the data $D$ are mutually independent:

$$p(D) = p\left(\{(x_1, c_1), \ldots, (x_n, c_n)\}\right) = \prod_{i=1}^{n} p(x_i, c_i)$$

(1) Usually *not* independent are any two events $X_i=x_i$ and $C_i=c_i$, $i=1, \ldots, n$:

$$p(x_i, c_i) \neq p(x_i) \cdot p(c_i)$$

For maximizing $p(D)$, see the maximum likelihood derivation of the **logistic loss** $L_\sigma(w)$.

- By **experiment design**, the probabilities, $p(x_i), i=1, \ldots, n$, are independent, i.e., the probability of the joint event $\{X_1=x_1, \ldots, X_n=x_n\}$ is equal to the product of the singleton events: $p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i)$.

A consistent and unbiased estimate for $p(x)$ is $\hat{p}(x) = |\{(x, \cdot) \in D\}| \cdot \frac{1}{|D|}$.

- By **experiment design**, the conditional probabilities, $p(c_i | x_i), i=1, \ldots, n$, are **invariant under covariate shift**, i.e., invariant under a change of $p(x_i)$. That is, the classification procedure, “determination of $c_i$ given some $x_i$”, always runs the same way, regardless of how often $x_i$ is encountered.
The invariance of $p(c_i \mid x_i)$ under a covariate shift can also be understood as the fact that any two events $X_i=x_i$ and $(C_i=c_i \mid X_i=x_i)$, $i = 1, \ldots, n$ are independent:

$$p(x, (c \mid x)) = p(x) \cdot p(c \mid x) = p(x, c)$$

However, this interpretation is problematic since standard probability theory does not allow a conditional event being combined with other events. See section Probability Basics of this part, conditional event algebra, and Lewis’s triviality result for details.

Within an outcome-only setting such as “flipping a coin”, the object features (coin diameter, coin age, etc.) are not used as predictors. I.e., one does not model the relationship between a response variable and predictors $x$ but models (the probability of) a sequence of outcomes $D = \{y_1, \ldots, y_n\}$ or $D = \{c_1, \ldots, c_n\}$.

The type of setting, be it predictor-response or outcome-only, is independent of data exploitation aspects such as

- discriminative versus generative,
- non-probabilistic versus probabilistic,
- maximum likelihood versus Bayes, or
- frequentist versus subjectivist.
Exploitation of Data

**Typical Learning Settings**

\[ D = \{(x_1, y_1), \ldots, (x_n, y_n)\}, \quad D = \{(x_1, c_1), \ldots, (x_n, c_n)\} \]

1. **RSS** (\(w\)):
   \[ \sum_{(x, y) \in D} (y - w^T x)^2 \]
   RSS for \(D\) under a linear model, parameterized by \(w\).
   Least squares estimate: \(\hat{w} = \arg\min_{w \in \mathbb{R}^{p+1}} \text{RSS}(w)\)

2. **\(p(D; w)\)**:
   \[ \prod_{(x, c) \in D} p(c \mid x; w) \]
   Probability of \(D\) under a logistic model, parameterized by \(w\).
   Maximum likelihood estimate:
   \(w_{ML} = \arg\max_{w \in \mathbb{R}^{p+1}} p(D; w)\)

3. **\(L(w)\)**:
   \[ \sum_{(x, c) \in D} l_{\sigma}(c, \sigma(w^T x)) \]
   Loss for \(D\) under a logistic model, parameterized by \(w\).
   Minimum loss (= maximum likelihood) estimate:
   \(\hat{w} = \arg\min_{w \in \mathbb{R}^{p+1}} L(w)\)

4. **\(p(c \mid x)\)**:
   \[ \frac{p(x \mid c) \cdot p(c)}{p(x)} \]
   Probability of \(c\) given \(x\) via Bayes’s rule.
   Maximum a posteriori class for \(x\):
   \(c_{\text{MAP}} = \arg\max_{c \in \{\oplus, \odot\}} p(c \mid x)\)

5. **\(p(D; \theta)\)**:
   \[ \binom{n}{k} \cdot \theta^k \cdot (1 - \theta)^{n-k} \]
   Probability of \(D\) under the binomial distribution, parameterized by \(\theta\).
   Maximum likelihood estimate:
   \(\theta_{ML} = \arg\max_{\theta \in [0;1]} p(D; \theta)\)

6. **\(p(\theta \mid D)\)**:
   \[ \frac{p(D \mid \theta) \cdot p(\theta)}{p(D)} \]
   Probability of \(\theta\) given \(D\) via Bayes’s rule.
   Maximum a posteriori hypothesis:
   \(\theta_{\text{MAP}} = \arg\max_{\theta \in \{\theta_1, \theta_2\}} p(\theta \mid D)\)
Exploitation of Data

Typical Learning Settings

\[ D = \{(x_1, y_1), \ldots, (x_n, y_n)\}, \quad D = \{(x_1, c_1), \ldots, (x_n, c_n)\} \]

(1) \( \text{RSS}(w) : \sum_{(x, y) \in D} (y - w^T x)^2 \) \hspace{1cm} \text{RSS for } D \text{ under a linear model, parameterized by } w. \text{ Least squares estimate: } \hat{w} = \arg\min_{w \in \mathbb{R}^{p+1}} \text{RSS}(w)

(2) \( p(D; w) : \prod_{(x, c) \in D} p(c | x; w) \) \hspace{1cm} \text{Probability of } D \text{ under a logistic model, parameterized by } w. \text{ Maximum likelihood estimate: } w_{ML} = \arg\max_{w \in \mathbb{R}^{p+1}} p(D; w)

(3) \( L(w) : \sum_{(x, c) \in D} l_{\sigma}(c, \sigma(w^T x)) \) \hspace{1cm} \text{Loss for } D \text{ under a logistic model, parameterized by } w. \text{ Minimum loss (= maximum likelihood) estimate: } \hat{w} = \arg\min_{w \in \mathbb{R}^{p+1}} L(w)

(4) \( p(c | x) : \frac{p(x | c) \cdot p(c)}{p(x)} \) \hspace{1cm} \text{Probability of } c \text{ given } x \text{ via Bayes's rule. Maximum a posteriori class for } x : c_{MAP} = \arg\max_{c \in \{\oplus, \ominus\}} p(c | x)

(5) \( p(D; \theta) : \binom{n}{k} \cdot \theta^k \cdot (1 - \theta)^{n-k} \) \hspace{1cm} \text{Probability of } D \text{ under the binomial distribution, parameterized by } \theta. \text{ Maximum likelihood estimate: } \theta_{ML} = \arg\max_{\theta \in [0; 1]} p(D; \theta)

(6) \( p(\theta | D) : \frac{p(D | \theta) \cdot p(\theta)}{p(D)} \) \hspace{1cm} \text{Probability of } \theta \text{ given } D \text{ via Bayes’s rule. Maximum a posteriori hypothesis: } \theta_{MAP} = \arg\max_{\theta \in \{\theta_1, \theta_2\}} p(\theta | D)
Exploitation of Data

Typical Learning Settings

\[ D = \{(x_1, y_1), \ldots, (x_n, y_n)\}, \quad D = \{(x_1, c_1), \ldots, (x_n, c_n)\} \]

(1) \( \text{RSS}(w) : \sum_{(x,y) \in D} (y - w^T x)^2 \)

Least squares estimate: \( \hat{w} = \arg\min_{w \in \mathbb{R}^{p+1}} \text{RSS}(w) \)

Probability of \( D \) under a logistic model, parameterized by \( w \). Maximum likelihood estimate:
\[ w_{\text{ML}} = \arg\max_{w \in \mathbb{R}^{p+1}} p(D; w) \]

Loss for \( D \) under a logistic model, parameterized by \( w \).
Minimum loss (= maximum likelihood) estimate:
\[ \hat{w} = \arg\min_{w \in \mathbb{R}^{p+1}} L(w) \]

Probability of \( c \) given \( x \) via Bayes's rule. Maximum a posteriori class for \( x \):
\[ c_{\text{MAP}} = \arg\max_{c \in \{\oplus, \ominus\}} p(c | x) \]

(2) \( p(D; w) : \prod_{(x,c) \in D} p(c | x; w) \)

(3) \( L(w) : \sum_{(x,c) \in D} l_{\sigma}(c, \sigma(w^T x)) \)

(4) \( p(c | x) : \frac{p(x | c) \cdot p(c)}{p(x)} \)

(5) \( p(D; \theta) : \binom{n}{k} \cdot \theta^k \cdot (1 - \theta)^{n-k} \)

Maximum likelihood estimate:
\[ \theta_{\text{ML}} = \arg\max_{\theta \in [0;1]} p(D; \theta) \]

Probability of \( D \) under the binomial distribution, parameterized by \( \theta \).

(6) \( p(\theta | D) : \frac{p(D | \theta) \cdot p(\theta)}{p(D)} \)

Maximum a posteriori hypothesis:
\[ \theta_{\text{MAP}} = \arg\max_{\theta \in \{\theta_1, \theta_2\}} p(\theta | D) \]
Exploitation of Data

Typical Learning Settings

\[ D = \{ (x_1, y_1), \ldots, (x_n, y_n) \}, \quad D = \{ (x_1, c_1), \ldots, (x_n, c_n) \} \]

1. RSS: \[ \text{RSS}(w) = \sum_{(x, y) \in D} (y - w^T x)^2 \]
   - RSS for \( D \) under a linear model, parameterized by \( w \).
   - Least squares estimate: \( \hat{w} = \arg\min_{w \in \mathbb{R}^{p+1}} \text{RSS}(w) \)

2. \( p(D; w) \): \[ \prod_{(x, c) \in D} p(c \mid x; w) \]
   - Probability of \( D \) under a logistic model, parameterized by \( w \).
   - Maximum likelihood estimate: \( w_{\text{ML}} = \arg\max_{w \in \mathbb{R}^{p+1}} p(D; w) \)

3. \( L(w) \): \[ \sum_{(x, c) \in D} l_{\sigma}(c, \sigma(w^T x)) \]
   - Loss for \( D \) under a logistic model, parameterized by \( w \).
   - Minimum loss (= maximum likelihood) estimate: \( \hat{w} = \arg\min_{w \in \mathbb{R}^{p+1}} L(w) \)

4. \( p(c \mid x) \): \[ \frac{p(x \mid c) \cdot p(c)}{p(x)} \]
   - Probability of \( c \) given \( x \) via Bayes’s rule.
   - Maximum a posteriori class for \( x \): \( c_{\text{MAP}} = \arg\max_{c \in \{\oplus, \ominus\}} p(c \mid x) \)

\[ D = \{ y_1, \ldots, y_n \}, \quad D = \{ c_1, \ldots, c_n \} \]

5. \( p(D; \theta) \): \[ \binom{n}{k} \cdot \theta^k \cdot (1 - \theta)^{n-k} \]
   - Probability of \( D \) under the binomial distribution, parameterized by \( \theta \).
   - Maximum likelihood estimate: \( \theta_{\text{ML}} = \arg\max_{\theta \in [0; 1]} p(D; \theta) \)

6. \( p(\theta \mid D) \): \[ \frac{p(D \mid \theta) \cdot p(\theta)}{p(D)} \]
   - Probability of \( \theta \) given \( D \) via Bayes’s rule.
   - Maximum a posteriori hypothesis: \( \theta_{\text{MAP}} = \arg\max_{\theta \in \{\theta_1, \theta_2\}} p(\theta \mid D) \)
Exploitation of Data

Typical Learning Settings

\[ D = \{(x_1, y_1), \ldots, (x_n, y_n)\}, \quad D = \{(x_1, c_1), \ldots, (x_n, c_n)\} \]

(1) RSS(\(w\)) : \[ \sum_{(x, y) \in D} (y - w^T x)^2 \]

RSS for \(D\) under a linear model, parameterized by \(w\). Least squares estimate: \(\hat{w} = \arg\min_{w \in \mathbb{R}^{p+1}} \text{RSS}(w)\)

(2) \(p(D; w) : \prod_{(x, c) \in D} p(c \mid x; w)\)

Probability of \(D\) under a logistic model, parameterized by \(w\). Maximum likelihood estimate:
\[ w_{\text{ML}} = \arg\max_{w \in \mathbb{R}^{p+1}} p(D; w) \]

(3) \(L(w) : \sum_{(x, c) \in D} l_\sigma(c, \sigma(w^T x))\)

Loss for \(D\) under a logistic model, parameterized by \(w\). Minimum loss (= maximum likelihood) estimate:
\[ \hat{w} = \arg\min_{w \in \mathbb{R}^{p+1}} L(w) \]

(4) \(p(c \mid x) : \frac{p(x \mid c) \cdot p(c)}{p(x)}\)

Probability of \(c\) given \(x\) via Bayes's rule. Maximum a posteriori class for \(x\) :
\[ c_{\text{MAP}} = \arg\max_{c \in \{\oplus, \ominus\}} p(c \mid x) \]

\(D = \{y_1, \ldots, y_n\}, \quad D = \{c_1, \ldots, c_n\} \)

(5) \(p(D; \theta) : {n \choose k} \cdot \theta^k \cdot (1 - \theta)^{n-k}\)

Probability of \(D\) under the binomial distribution, parameterized by \(\theta\). Maximum likelihood estimate:
\[ \theta_{\text{ML}} = \arg\max_{\theta \in [0; 1]} p(D; \theta) \]

(6) \(p(\theta \mid D) : \frac{p(D \mid \theta) \cdot p(\theta)}{p(D)}\)

Probability of \(\theta\) given \(D\) via Bayes's rule. Maximum a posteriori hypothesis:
\[ \theta_{\text{MAP}} = \arg\max_{\theta \in \{\theta_1, \theta_2\}} p(\theta \mid D) \]
Exploitation of Data

Typical Learning Settings

\[ D = \{(x_1, y_1), \ldots, (x_n, y_n)\}, \quad D = \{(x_1, c_1), \ldots, (x_n, c_n)\} \]

1. **RSS** (\(w\)): \[ \sum_{(x,y) \in D} (y - w^T x)^2 \]
   
   RSS for \(D\) under a linear model, parameterized by \(w\).
   
   Least squares estimate: \(\hat{w} = \arg\min_{w \in \mathbb{R}^{p+1}} \text{RSS}(w)\)

2. **\(p(D; w)\)**: \[ \prod_{(x,c) \in D} p(c \mid x; w) \]
   
   Probability of \(D\) under a logistic model, parameterized by \(w\).
   
   Maximum likelihood estimate: \(w_{ML} = \arg\max_{w \in \mathbb{R}^{p+1}} p(D; w)\)

3. **\(L(w)\)**: \[ \sum_{(x,c) \in D} l_\sigma(c, \sigma(w^T x)) \]
   
   Loss for \(D\) under a logistic model, parameterized by \(w\).
   
   Minimum loss (= maximum likelihood) estimate: \(\hat{w} = \arg\min_{w \in \mathbb{R}^{p+1}} L(w)\)

4. **\(p(c \mid x)\)**: \[ \frac{p(x \mid c) \cdot p(c)}{p(x)} \]
   
   Probability of \(c\) given \(x\) via Bayes's rule. Maximum a posteriori class for \(x\): \(c_{MAP} = \arg\max_{c \in \{\oplus, \odot\}} p(c \mid x)\)

5. **\(p(D; \theta)\)**: \[ \binom{n}{k} \cdot \theta^k \cdot (1 - \theta)^{n-k} \]
   
   Probability of \(D\) under the binomial distribution, parameterized by \(\theta\).
   
   Maximum likelihood estimate: \(\theta_{ML} = \arg\max_{\theta \in [0;1]} p(D; \theta)\)

6. **\(p(\theta \mid D)\)**: \[ \frac{p(D \mid \theta) \cdot p(\theta)}{p(D)} \]
   
   Probability of \(\theta\) given \(D\) via Bayes's rule. Maximum a posteriori hypothesis: \(\theta_{MAP} = \arg\max_{\theta \in \{\theta_1, \theta_2\}} p(\theta \mid D)\)
Exploitation of Data

Typical Learning Settings

\[ D = \{ (x_1, y_1), \ldots, (x_n, y_n) \}, \quad D = \{ (x_1, c_1), \ldots, (x_n, c_n) \} \]

(1) \( \text{RSS}(w) : \sum_{(x,y) \in D} (y - w^T x)^2 \) \quad \text{RSS for } D \text{ under a linear model, parameterized by } w. \quad \text{Least squares estimate: } \hat{w} = \arg\min_{w \in \mathbb{R}^{p+1}} \text{RSS}(w)

(2) \( p(D; w) : \prod_{(x,c) \in D} p(c \mid x; w) \) \quad \text{Probability of } D \text{ under a logistic model, parameterized by } w. \quad \text{Maximum likelihood estimate: } \hat{w}_{\text{ML}} = \arg\max_{w \in \mathbb{R}^{p+1}} p(D; w)

(3) \( L(w) : \sum_{(x,c) \in D} l_\sigma(c, \sigma(w^T x)) \) \quad \text{Loss for } D \text{ under a logistic model, parameterized by } w. \quad \text{Minimum loss (= maximum likelihood) estimate: } \hat{w} = \arg\min_{w \in \mathbb{R}^{p+1}} L(w)

(4) \( p(c \mid x) : \frac{p(x \mid c) \cdot p(c)}{p(x)} \) \quad \text{Probability of } c \text{ given } x \text{ via Bayes’s rule. Maximum a posteriori class for } x : \hat{c}_{\text{MAP}} = \arg\max_{c \in \{\oplus, \ominus\}} p(c \mid x)

(5) \( p(D; \theta) : \binom{n}{k} \cdot \theta^k \cdot (1 - \theta)^{n-k} \) \quad \text{Probability of } D \text{ under the binomial distribution, parameterized by } \theta. \quad \text{Maximum likelihood estimate: } \hat{\theta}_{\text{ML}} = \arg\max_{\theta \in [0;1]} p(D; \theta)

(6) \( p(\theta \mid D) : \frac{p(D \mid \theta) \cdot p(\theta)}{p(D)} \) \quad \text{Probability of } \theta \text{ given } D \text{ via Bayes’s rule. Maximum a posteriori hypothesis: } \hat{\theta}_{\text{MAP}} = \arg\max_{\theta \in \{\theta_1, \theta_2\}} p(\theta \mid D)
Remarks (predictor-response vs. outcome-only setting):

1),...,4) Predictor-response setting, \( x \rightarrow y \) or \( x \rightarrow c \). The relation between \( x \) and \( y \) or \( c \) is captured by a model function \( y(x) \). The data \( D \) is exploited to fit \( y(x) \), which in turn means to determine a parameter \( w \) or parameter vector \( w \) for \( y(x) \). Modeling and predicting a quantitative response variable \( y \) is a regression task; modeling and predicting a categorical response variable \( c \) is a classification task.

An example for a categorical predictor-response setting is the classification of an email as spam \((c = \oplus)\) or ham \((c = \ominus)\), given a vector \( x \) of linguistic features for that email.

5), 6) Outcome-only setting, \( y_1, \ldots, y_n \) or \( c_1, \ldots, c_n \). Modeling a sole outcome variable means to fit the data \( D \) using a suited distribution function, which in turn means to determine the distribution parameter \( \theta \) or distribution parameters \( \theta \). Again, one can distinguish between different measurement scales, such as quantitative \((y)\) or categorical \((c)\).

An example for a categorical outcome-only setting is a coin flip experiment where one has to fit the observations (number of heads and tails) under the binomial distribution, which in turn means to determine the distribution parameter \( \theta \).

1),...,6) Depending on the experiment setting, i.e., fitting of a model function vs. fitting of a distribution, either the symbol \( w \) (or \( w \)), or the symbol \( \theta \) (or \( \theta \)) may be used to denote the parameter (or parameter vector).
Remarks (discriminative vs. generative approach):

(1), (2), (3) Discriminative approach to classification. Exploit the data to determine a decision boundary. Typically, “discriminative” implies “frequentist”.

The optimization (argmin, argmax) considers \( p(x) \), the distribution of the independent variables \( x \), implicitly via the multiplicity of \( x \) in the data \( D \). Recall that \( D \) is a multiset of examples.

(2), (3), (5) Maximum likelihood (ML) principle to parameter estimation.

(2) Recall the identities from the maximum likelihood derivation of the \textit{logistic loss} \( L_{\sigma}(w) \):

\[
p(D; w) = \prod_{(x,c) \in D} p(x, c; w), \quad \argmax_{w \in \mathbb{R}^{p+1}} p(D; w) = \argmax_{w \in \mathbb{R}^{p+1}} \prod_{(x,c) \in D} p(c | x; w)
\]

(1), (2) If the data comes from an exponential family and mild conditions are satisfied, least-squares estimates and maximum-likelihood estimates are identical.

(2), (3) Probabilistic model. The conditional class probability function (CCPF), \( p(c | x) \), is estimated for all feature vectors (= at all quantiles). The model is not generative since the distribution of the independent variable, \( p(x) \), is not modeled (but of course exploited implicitly via \( D \)).

Maximizing the probability under a logistic model is equivalent to minimizing the logistic loss \( L_{\sigma} \). Hence, \( w_{\text{ML}} = \hat{w} \).
Remarks (discriminative vs. generative approach) : (continued)

(4) Generative approach to classification. Exploit the data \( D \) (here: estimate \( p(x \mid c) \) and \( p(c) \) for all \( x \) and \( c \)) to provide a model for the joint probability distribution, \( p(x, c) \), from which \( D \) is sampled.

(5) Generative approach. Assuming the conditions of the binomial data model, exploit the data \( D \) (here: estimate the parameter \( \theta \)) to provide a model for the binomial probability distribution, \( p(c) \), from which \( D \) is sampled.

(6) Generative or discriminative approach. \( p(\theta \mid D) \) can be estimated by either providing (\( \rightarrow \) generative) or by \textit{not} providing (\( \rightarrow \) discriminative) a model for the probability distribution from which \( D \) is sampled.
Remarks (ML principle vs. Bayes method):

\[ (1), (2), (3) \]
\[ (5) \]

\( w \) (as well as \( \theta \)) is not the realization of a random variable—which would come along with a distribution—but an **exogenous parameter**, which is varied in order to find the maximum probability \( p(D; w) \) (or \( p(D; \theta) \)) or the minimum loss \( L(w) \).

The fact that \( w \) (or \( \theta \)) is an exogenous parameter and not a realization of a random variable is reflected by the notation, which uses a »;« instead of a »|« in the argument of \( p() \).

(4) Application of Bayes’s rule, presupposing that one can estimate the likelihoods \( p(x \mid \cdot) \) (\( p(x_j \mid \cdot) \) in case of Naive Bayes) at higher fidelity than the conditional class probabilities, \( p(\cdot \mid x) \), from the data.

Under the Naive Bayes Assumption, \( p(x \mid c) \) is modeled as \( \prod_{j=1}^{p} p(x_j \mid c) \).

(4), (6) Likelihoods, \( p(x \mid \cdot) \), \( p(D \mid \cdot) \), are computed for events under alternative classes \( c \) or parameters \( \theta \). The settings differ in that an event in (4) is about a feature vector \( x \), while an event in (6) is about a sequence \( D \). (4) may (but not need to) apply the Naive Bayes assumption to compute the likelihood \( p(x \mid c) \), which is a common approximation for a nominal feature space and if data are sparse. For (6), if the data originate from a coin flip experiment, the likelihood \( p(D \mid \theta) \) is computed via the binomial distribution.

If the prior probabilities, \( p(c) \) or \( p(\theta) \), are estimated also from \( D \), we follow the frequentist paradigm; if the priors rely on subjective assessments we follow the subjectivist paradigm.

If we assume uniform priors, i.e., the \( p(c) \) or the \( p(\theta) \) are equally probable, MAP estimates and ML estimates are equal since \( p(c \mid x) \propto p(x \mid c) \) or \( p(\theta \mid D) \propto p(D \mid \theta) \), where »\( \propto \)« means “is proportional to”.

ML:VII-114 Bayesian Learning © STEIN/VÖLSKE 2023
Exploitation of Data

Learning Approaches Overview

- **Data exploitation**
  - **non-probabilistic**
    - discriminative (→ frequentist)
    - probabilistic
      - frequentist
      - subjectivist
  - generative (→ probabilistic)
    - **Support vector machine**
    - (1) Linear regression with least square estimates from $D$
    - (2) Logistic regression via $p()$ with ML estimates from $D$
    - (3) Logistic regression via $L()$ with ML estimates from $D$
    - (4) Bayes with ML estimates from $D$ as priors
      - (5) Probability model with ML estimate from $D$
      - (6) Bayes with subjective priors

$D = \{(x_1,y_1),\ldots,(x_n,y_n)\}$, $D = \{(x_1,c_1),\ldots,(x_n,c_n)\}$

$D = \{y_1,\ldots,y_n\}$, $D = \{c_1,\ldots,c_n\}$
Exploitation of Data
Learning Approaches Overview (continued)

data exploitation

- **discriminative**
  - frequentist
  - non-probabilistic

- **probabilistic**
  - generative
  - (→ probabilistic)
  - frequentist
  - subjective

- **generative**
  - (→ probabilistic)

- **subjectivist**

Support vector machine

1. Linear regression with least square estimates from $D$
2. Logistic regression via $p()$ with ML estimates from $D$
3. Logistic regression via $L()$ with ML estimates from $D$
4. Bayes with ML estimates from $D$ as priors
5. Probability model with ML estimate from $D$
6. Bayes with subjective priors

$D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, $D = \{(x_1, c_1), \ldots, (x_n, c_n)\}$
$D = \{y_1, \ldots, y_n\}$, $D = \{c_1, \ldots, c_n\}$

**Discriminative:** Determine a boundary to split $D$. → No model for the distribution of $D$.

**Generative:** Provide a model for the probability distribution from which $D$ is sampled.
Exploitation of Data
Learning Approaches Overview (continued)

- **non-probabilistic**: Threshold some model function (typically at zero). → Classification, Labeling
- **probabilistic**: Estimate $p(c \mid x)$ at all quantiles. → Class probability estimation, CCPF

**discriminative** (→ frequentist)
- Support vector machine
  - (1) Linear regression with least square estimates from $D$
  - (2) Logistic regression via $p()$ with ML estimates from $D$
  - (3) Logistic regression via $L()$ with ML estimates from $D$
  - (4) Bayes with ML estimates from $D$ as priors
  - (5) Probability model with ML estimate from $D$
  - (6) Bayes with subjective priors

**generative** (→ probabilistic)
- $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, $D = \{(x_1, c_1), \ldots, (x_n, c_n)\}$
- $D = \{y_1, \ldots, y_n\}$, $D = \{c_1, \ldots, c_n\}$

ML:VII-117 Bayesian Learning © STEIN/VÖLSKE 2023
Exploitation of Data
Learning Approaches Overview (continued)

- **non-probabilistic**
  - discriminative
    - (→ frequentist)
  - probabilistic
    - frequentist
    - (→ probabilistic)
    - subjectivist

- **probabilistic**
  - discriminative
    - (→ frequentist)
  - generative
    - (→ probabilistic)

**Support vector machine**

1. Linear regression with least square estimates from $D$
2. Logistic regression via $p()$ with ML estimates from $D$
3. Logistic regression via $L()$ with ML estimates from $D$
4. Bayes with ML estimates from $D$ as priors
5. Probability model with ML estimate from $D$
6. Bayes with subjective priors

**ML:VII-118 Bayesian Learning © STEIN/VÖLSKE 2023**

**frequentist**: Consider a unique mechanism that generated the data $D$.

**subjectivist**: Specify beliefs for alternative mechanisms one of which generated $D$. 
Remarks:

- We call a data exploitation approach “generative” if it provides us with a model for the probability distribution from which $D$ is sampled. With such a model we are able to generate arbitrary samples from the population where $D$ is sampled from.

- The overview does not show all but common combinations. In particular:
  - Typically, “discriminative” implies “frequentist”. The inverse does not apply: consider a Bayes classifier with priors estimated from the data.
  - Typically, “generative” implies “probabilistic”. The inverse does not apply: logistic regression provides a probabilistic model to classification.

- Discriminative approaches are further distinguished as “non-probabilistic” or “probabilistic”.

- Generative approaches are further distinguished as “frequentist” or “subjectivist”.
Chapter ML:VII (continued)

VII. Bayesian Learning

- Approaches to Probability
- Conditional Probability
- Bayes Classifier
- Exploitation of Data
- Frequentist versus Subjectivist
Frequentist versus Subjectivist
Logistic Regression versus Naive Bayes

- Frequentist: (\rightarrow \text{frequentist})
  - Discriminative
  - Probabilistic
- Subjectivist: (\rightarrow \text{probabilistic})
  - Generative
  - Non-probabilistic

Support vector machine
1. Linear regression with least square estimates from \(D\)
2. Logistic regression via \(p()\) with ML estimates from \(D\)
3. Logistic regression via \(L()\) with ML estimates from \(D\)
4. Bayes with ML estimates from \(D\) as priors
5. Probability model with ML estimate from \(D\)
6. Bayes with subjective priors

Data exploitation examples:
- \(D = \{(x_1, y_1), \ldots, (x_n, y_n)\}\)
- \(D = \{(x_1, c_1), \ldots, (x_n, c_n)\}\)
- \(D = \{y_1, \ldots, y_n\}\)
- \(D = \{c_1, \ldots, c_n\}\)
Frequentist versus Subjectivist
Logistic Regression versus Naive Bayes

(2) \( \mathbf{w}_{\text{ML}} = \arg\max_{\mathbf{w} \in \mathbb{R}^{p+1}} \prod_{(x,c) \in D} p(c \mid x; \mathbf{w}) \) (logistic regression)

(4) \( c_{\text{MAP}} = \arg\max_{c \in \{\oplus, \ominus\}} p(c \mid x) \)

Observation 1. Both approaches maximize \( p(D) \):

- (2), the "ML principle", determines the parameters \( \mathbf{w} \) of the logistic model function such that \( \prod_{D} p(c \mid x) \) becomes maximum. Note that a parameter vector \( \mathbf{w} \) that maximizes \( \prod_{D} p(c \mid x) \) will also maximize \( \prod_{D} p(x, c) \), and thus \( p(D) \) (under the i.i.d. assumption).

- (4), the "Bayes method", determines for a given \( x \) its most probable class. By choosing \( c_{\text{MAP}} \) for each \( x \), Bayes maximizes \( p(D) \) by maximizing each factor of \( \prod_{D} p(c \mid x) \). Note that \( p(x) \) is constant per factor. Recall that Naive Bayes approximates \( p(x \mid c) \) with \( \prod_{j=1}^{p} p(x_j \mid c) \).
Frequentist versus Subjectivist
Logistic Regression versus Naive Bayes

(2) \( w_{ML} = \arg\max_{w \in \mathbb{R}^{p+1}} \prod_{(x,c) \in D} p(c \mid x; w) \) \hspace{1cm} \text{(logistic regression)}

(4) \( c_{MAP} = \arg\max_{c \in \{\oplus, \ominus\}} \frac{p(x \mid c) \cdot p(c)}{p(x)} \) \hspace{1cm} \text{(Bayes)}

Observation 1. Both approaches maximize \( p(D) \):

- (2), the “ML principle”, determines the parameters \( w \) of the logistic model function such that \( \prod_{D} p(c \mid x) \) becomes maximum. Note that a parameter vector \( w \) that maximizes \( \prod_{D} p(c \mid x) \) will also maximize \( \prod_{D} p(x, c) \), and thus \( p(D) \) (under the i.i.d. assumption).

- (4), the “Bayes method”, determines for a given \( x \) its most probable class. By choosing \( c_{MAP} \) for each \( x \), Bayes maximizes \( p(D) \) by maximizing each factor of \( \prod_{D} p(c \mid x) \). Note that \( p(x) \) is constant per factor. Recall that Naive Bayes approximates \( p(x \mid c) \) with \( \prod_{j=1}^{p} p(x_j \mid c) \).
Frequentist versus Subjectivist
Logistic Regression versus Naive Bayes

(2) $\mathbf{w}_{\text{ML}} = \arg\max_{\mathbf{w} \in \mathbb{R}^{p+1}} \prod_{(x,c) \in D} p(c \mid x; \mathbf{w})$ \hspace{1cm} \text{(logistic regression)}

(4) $c_{\text{MAP}} = \arg\max_{c \in \{\oplus, \ominus\}} \frac{\prod_{j=1}^{p} p(x_j \mid c) \cdot p(c)}{p(x)}$ \hspace{1cm} \text{(Naive Bayes)}

Observation 1. Both approaches maximize $p(D)$:

- (2), the “ML principle”, determines the parameters $\mathbf{w}$ of the logistic model function such that $\prod_{D} p(c \mid x)$ becomes maximum. Note that a parameter vector $\mathbf{w}$ that maximizes $\prod_{D} p(c \mid x)$ will also maximize $\prod_{D} p(x, c)$, and thus $p(D)$ (under the i.i.d. assumption).

- (4), the “Bayes method”, determines for a given $x$ its most probable class. By choosing $c_{\text{MAP}}$ for each $x$, Bayes maximizes $p(D)$ by maximizing each factor of $\prod_{D} p(c \mid x)$. Note that $p(x)$ is constant per factor. Recall that Naive Bayes approximates $p(x \mid c)$ with $\prod_{j=1}^{p} p(x_j \mid c)$. 
Frequentist versus Subjectivist
Logistic Regression versus Naive Bayes

(2) \[ w_{ML} = \arg\max_{w \in \mathbb{R}^{p+1}} \prod_{(x,c) \in D} p(c \mid x; w) \] (logistic regression)

(4) \[ c_{MAP} = \arg\max_{c \in \{\oplus, \ominus\}} \prod_{j=1}^{p} p(x_j \mid c) \cdot p(c) \] (Naive Bayes)

Observation 1. Both approaches maximize \( p(D) \):

- (2), the “ML principle”, determines the parameters \( w \) of the logistic model function such that \( \prod_D p(c \mid x) \) becomes maximum. Note that a parameter vector \( w \) that maximizes \( \prod_D p(c \mid x) \) will also maximize \( \prod_D p(x, c) \), and thus \( p(D) \) (under the i.i.d. assumption).

- (4), the “Bayes method”, determines for a given \( x \) its most probable class. By choosing \( c_{MAP} \) for each \( x \), Bayes maximizes \( p(D) \) by maximizing each factor of \( \prod_D p(c \mid x) \). Note that \( p(x) \) is constant per factor. Recall that Naive Bayes approximates \( p(x \mid c) \) with \( \prod_{j=1}^{p} p(x_j \mid c) \).
Frequentist versus Subjectivist
Logistic Regression versus Naive Bayes  [data exploitation examples]  (continued)

(2) $w_{ML} = \arg\max_{w \in \mathbb{R}^{p+1}} \prod_{(x,c) \in D} p(c \mid x; w)$ (logistic regression)

(4) $c_{MAP} = \arg\max_{c \in \{ \oplus, \ominus \}} \prod_{j=1}^{p} p(x_j \mid c) \cdot p(c)$ (Naive Bayes)

Observation 1. Both approaches maximize $p(D)$:

- (2), the “ML principle”, determines the parameters $w$ of the logistic model function such that $\prod_D p(c \mid x)$ becomes maximum. Note that a parameter vector $w$ that maximizes $\prod_D p(c \mid x)$ will also maximize $\prod_D p(x, c)$, and thus $p(D)$ (under the i.i.d. assumption).

- (4), the “Bayes method”, determines for a given $x$ its most probable class. By choosing $c_{MAP}$ for each $x$, Bayes maximizes $p(D)$ by maximizing each factor of $\prod_D p(c \mid x)$. Note that $p(x)$ is constant per factor. Recall that Naive Bayes approximates $p(x \mid c)$ with $\prod_{j=1}^{p} p(x_j \mid c)$. 
Frequentist versus Subjectivist
Logistic Regression versus Naive Bayes

\[(2) \quad w_{ML} = \arg\max_{w \in \mathbb{R}^{p+1}} \prod_{(x, c) \in D} p(c \mid x; w) \quad \text{(logistic regression)}\]

\[(4) \quad c_{MAP} = \arg\max_{c \in \{\ominus, \oplus\}} \prod_{j=1}^{p} p(x_j \mid c) \cdot p(c) \quad \text{(Naive Bayes)}\]

Observation 2 (corollary). Both approaches model the covariate distribution:

- (2), the “ML principle”, considers \(p(x)\), the distribution of the independent variables \(x\), implicitly via the multiplicity of \(x\) in the data \(D\). Recall that \(D\) is a multiset of examples.

- (4), the “Bayes method”, as a generative approach, models \(p(x \mid c)\) and \(p(c)\), and hence also \(p(x, c), p(x),\) and \(p(c \mid x)\). The likelihoods, \(p(x \mid c)\) (or \(p(x_j \mid c)\) under Naive Bayes), are estimated from \(D\); the priors, \(p(c)\), may be estimated by subjective assessments.
Both approaches maximize $p(D)$ by maximizing $\prod_D p(c \mid x)$. Estimating $p(c \mid x)$ is usually significantly easier than estimating $p(x, c)$.

(4) Naive Bayes models $p(x \mid c)$ as $\prod_{j=1}^p p(x_j \mid c)$, where $p(x_j \mid c)$ is estimated as $\hat{p}(x_j \mid c)$, $\hat{p}(x_j \mid c) = |\{(x, c) \in D : x_j = x_j\}| / |\{(\cdot, c) \in D\}|$.

Similarly, $p(c)$ can be estimated as $\hat{p}(c)$, $\hat{p}(c) = |\{c \in D\}|$; but, also a dedicated (and subjective) prior probability model can be stated.

$p(x)$ can be computed with the Law of Total Probability, $p(x) = \sum_{c \in \{\oplus, \ominus\}} p(x \mid c) \cdot p(c)$. Note, however, that $p(x)$ is not required to compute $c_{\text{MAP}}$ for $x$.

(4) If for the Bayes method—aside from the likelihoods $p(x_j \mid c)$—also the class priors, $p(c)$, are computed from $D$, we follow the frequentist paradigm, similar to the ML principle. Only if the values for $p(c)$ (= the prior probability model) rely on subjective assessments, the Bayes method can be considered as subjectivist.

Whether to apply the ML principle or the Bayes method is not a free choice; it depends on

- the availability of data $D$,
- the conditional strengths of the likelihoods, $p(x \mid c)$,
- the reliability of the assessments for the prior probabilities, $p(c)$, and,
- whether or not subjective assessments shall be considered to estimate the priors $p(c)$.

Synonymous: covariate, independent, predictor variable / distribution.
Remarks: (continued)

- Observe the subtle distinction between “Bayes rule” and “Bayes method” made here. With the former we refer to the identity that connects the posterior probability, $P(A \mid B)$, and the likelihood, $P(B \mid A)$ (the “reversal of condition and consequence”). With the latter we refer to the *parameter estimation principle* where the maximum a posteriori probability is determined.

- Note that a class-conditional event “$X=x \mid C=c$” does not necessarily model a cause-effect relation: the event “$C=c$” may cause—but does not need to cause—the event “$X=x$”.

Examples:
- A disease $c$ will cause the symptoms $x$ (but not vice versa).
- Weather conditions $x$ will cause the decision “$\text{EnjoySurfing}=\text{yes}$” (but not vice versa).

Similarly, also if $x$ is the independent variable of a function $y(x)$ that maps features to classes $c$, the cause-effect direction is not necessarily $x \rightarrow c$, but can also be the other way around: Consider $y(x) = c$ with “disease $c$” $\rightarrow$ ”symptoms $x$“.
Frequentist versus Subjectivist
Logistic Regression versus Naive Bayes: Example

A multiset of examples $D$:

<table>
<thead>
<tr>
<th></th>
<th>URLs</th>
<th>Spelling errors</th>
<th>Spam</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5</td>
<td>3</td>
<td>yes</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>1</td>
<td>no</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>3</td>
<td>yes</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>0</td>
<td>no</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>0</td>
<td>yes</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>15</td>
<td>1</td>
<td>4</td>
<td>no</td>
</tr>
<tr>
<td>16</td>
<td>1</td>
<td>4</td>
<td>yes</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>20</td>
<td>0</td>
<td>4</td>
<td>no</td>
</tr>
</tbody>
</table>

Learning task:

- Fit $D$ to compute a classifier for feature vectors $x$, $x \notin D$. 
Frequentist versus Subjectivist
Logistic Regression versus Naive Bayes: Example  (continued)

A multiset of examples $D$:

<table>
<thead>
<tr>
<th>URLs</th>
<th>Spelling errors</th>
<th>Spam</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>15</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>16</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>20</td>
<td>0</td>
<td>4</td>
</tr>
</tbody>
</table>

Learning task:
- Fit $D$ to compute a classifier for feature vectors $x$, $x \notin D$. 
Frequentist versus Subjectivist
Logistic Regression versus Naive Bayes: Example  (continued)

A multiset of examples $D$:

<table>
<thead>
<tr>
<th></th>
<th>URLs</th>
<th>Spelling errors</th>
<th>Spam</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5</td>
<td>3</td>
<td>yes</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>1</td>
<td>no</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>3</td>
<td>yes</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>0</td>
<td>no</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>0</td>
<td>yes</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>15</td>
<td>1</td>
<td>4</td>
<td>no</td>
</tr>
<tr>
<td>16</td>
<td>1</td>
<td>4</td>
<td>yes</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>20</td>
<td>0</td>
<td>4</td>
<td>no</td>
</tr>
</tbody>
</table>

Learning task:

- Fit $D$ to compute a classifier for feature vectors $x$, $x \notin D$. 
Frequentist versus Subjectivist
Logistic Regression versus Naive Bayes: Conditional Class Probabilities

Logistic regression:

Distribution of $D$. 

ML: VII-133 Bayesian Learning © STEIN/VÖLSKE 2023
Logistic regression:

- Hyperplane $\langle w_{ML}, x \rangle = 0$. 

![Graph showing logistic regression with variables X1: Spelling errors and X2: URLs.](image-url)
Frequentist versus Subjectivist
Logistic Regression versus Naive Bayes: Conditional Class Probabilities (continued)

Logistic regression:

- Conditional class probabilities computed with $w_{ML}$, the ML estimate for $w$ given $D$. 
Frequentist versus Subjectivist
Logistic Regression versus Naive Bayes: Conditional Class Probabilities (continued)

Logistic regression:

- $X_1$: Spelling errors
- $X_2$: URLs

Training error.

$D$ is the decision boundary.

$p(\cdot | x) = 1.0$
Frequentist versus Subjectivist
Logistic Regression versus Naive Bayes: Conditional Class Probabilities (continued)

Logistic regression:

- Training error.

Naive Bayes:

- Distribution of $D$. 

$X_1$: Spelling errors
$X_2$: URLs

$p(\bullet|x) = 1.0$
$p(\bullet|x) = 1.0$

Hyperplane distance

Ham
Spam

ML:VII-137 Bayesian Learning © STEIN/VÖLSKE 2023
Frequentist versus Subjectivist

Logistic Regression versus Naive Bayes: Conditional Class Probabilities (continued)

Logistic regression:

- Training error.

Naive Bayes:

- Conditional class probabilities computed for the respective MAP class, using $p(c)$ estimates from $D$. 

ML:VII-138 Bayesian Learning © STEIN/VÖLSKE 2023
Frequentist versus Subjectivist
Logistic Regression versus Naive Bayes: Conditional Class Probabilities (continued)

Logistic regression:

Naive Bayes:

- Training error.
- Training error.
Logistic regression:  

- Computation of a hyperplane.  
- Approach: minimization of accumulated “misclassification distances” for examples in $D$.  
- Discriminative and probabilistic.

Naive Bayes:  

- Computation of a probability distribution.  
- Basis: class-conditional feature and class frequencies in $D$.  
- Generative (implies probabilistic).
Remarks:

- Both approaches, logistic regression and Naive Bayes, estimate the conditional class probability function, \( p(\text{Spam} \mid x) \) or \( p(\text{Ham} \mid x) = 1 - p(\text{Spam} \mid x) \). However, the two estimation approaches follow very different concepts.

- Generalization characteristic:
  - The conditional class probability function as computed via logistic regression decides not only the feature space \( \{0, 1, 2, 3, 4, 5\}^2 \) but the entire \( \mathbb{R}^2 \). (whether this makes sense is another question)
  - The conditional class probability function as computed via Naive Bayes provides class probability estimates for \( x \in \{0, 1, 2, 3, 4, 5\}^2 \). The probabilities are estimated from the class-conditional feature frequencies (likelihood estimates) and class frequencies, \( \hat{p}(x_1 \mid c) \), \( \hat{p}(x_2 \mid c) \), and \( \hat{p}(c) \), as found in \( D \). Note that a vector \( x = (x_1, x_2)^T \) gets the probability of zero for class \( c \), if \( x_1 \) or \( x_2 \) does not occur in some feature vector with class label \( c \) in \( D \).

- Handling of class imbalance and covariate distribution:
  - Logistic regression considers the \( p(c) \) and the \( p(x) \) implicitly via their multiplicity in \( D \). I.e., the learned parameter vector \( w \) has the class imbalance as well as the covariate distribution “compiled in”.
  - Naive Bayes, again, estimates the \( p(c) \) and the \( p(x) \) from the frequencies in \( D \). More specifically, \( p(x) \) can be estimated from \( \hat{p}(x_1 \mid c) \), \( \hat{p}(x_2 \mid c) \), and \( \hat{p}(c) \) with the Law of Total Probability. Note that the computation of \( p(x) \) is not necessary for a ranking (= classification without class membership probability).
Frequentist versus Subjectivist
Naive Bayes: Smoothing and Continuous Likelihoods
Frequentist versus Subjectivist

Naive Bayes: Prior Probability Models

Comparison of the conditional class probability function, \( p(c \mid x) \), under Naive Bayes for three different prior probability models (\( = \) assessments of class priors), \( p(c) \).

\( p(c) \) estimates from \( D \)

\[
\begin{align*}
P_a(C=\text{Spam}) &= \hat{p}(\text{Spam}) = 0.45 \\
P_a(C=\text{Ham}) &= \hat{p}(\text{Ham}) = 0.55
\end{align*}
\]

Subjective assessments for \( p(c) \)

\[
\begin{align*}
P_b(C=\text{Spam}) &= 0.6 \\
P_b(C=\text{Ham}) &= 0.4 \\
P_c(C=\text{Spam}) &= 0.8 \\
P_c(C=\text{Ham}) &= 0.2
\end{align*}
\]
Frequentist versus Subjectivist
Classification: Bayes Optimum versus MAP versus Ensemble
Recall the Bayes rule,

\[ P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}, \]

with \( A \) and \( B \) in the role of a “hypothesis event”, \( H=h \), and a “data event”, \( D=D \),

\[ P(H=h \mid D=D) = \frac{P(D=D \mid H=h) \cdot P(H=h)}{P(D=D)} \]
Recall the Bayes rule,

\[ P(A | B) = \frac{P(B | A) \cdot P(A)}{P(B)} \]

with \( A \) and \( B \) in the role of a “hypothesis event”, \( H=h \), and a “data event”, \( D=D \),

\[ P(H=h | D=D) = \frac{P(D=D | H=h) \cdot P(H=h)}{P(D=D)} \]

rewritten using probability mass functions, pmf, (in case of discrete events):

\[ p(h | D) = \frac{p(D | h) \cdot p(h)}{p(D)} \]

- Likelihood: How well does \( h \) explain (= entail, induce, evoke) the data \( D \)?
- Prior: How probable is the hypothesis \( h \) a priori (= in principle)?
- Normalization: How probable is the observation of the data \( D \)?
- Posterior: How probable is the hypothesis \( h \) when observing the data \( D \)?
Frequentist versus Subjectivist
Advanced Bayesian Decision Making (continued)

Recall the Bayes rule,

\[ P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)} \]

with \( A \) and \( B \) in the role of a “hypothesis event”, \( H=\hat{h} \), and a “data event”, \( D=D \),

\[ P(\hat{H}=\hat{h} \mid D=D) = \frac{P(D=D \mid \hat{H}=\hat{h}) \cdot P(\hat{H}=\hat{h})}{P(D=D)} \]

rewritten using probability mass functions, pmf, (in case of discrete events):

\[ p(\hat{h} \mid D) = \frac{p(D \mid \hat{h}) \cdot p(\hat{h})}{p(D)} \]

- Likelihood: How well does \( \hat{h} \) explain (= entail, induce, evoke) the data \( D \)?
- Prior: How probable is the hypothesis \( \hat{h} \) a priori (= in principle)?
- Normalization: How probable is the observation of the data \( D \)?
- Posterior: How probable is the hypothesis \( \hat{h} \) when observing the data \( D \)?
Recall the Bayes rule,

\[ P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)} , \]

with \( A \) and \( B \) in the role of a “hypothesis event”, \( H=h \), and a “data event”, \( D=D \),

\[ P(H=h \mid D=D) = \frac{P(D=D \mid H=h) \cdot P(H=h)}{P(D=D)} \]

rewritten using probability mass functions, pmf, (in case of discrete events):

\[ p(h \mid D) = \frac{p(D \mid h) \cdot p(h)}{p(D)} \]

- Likelihood: How well does \( h \) explain (= entail, induce, evoke) the data \( D \)?
- Prior: How probable is the hypothesis \( h \) a priori (= in principle)?
- Normalization: How probable is the observation of the data \( D \)?
- Posterior: How probable is the hypothesis \( h \) when observing the data \( D \)?
Recall the Bayes rule,

\[ P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}, \]

with \( A \) and \( B \) in the role of a “hypothesis event”, \( H=h \), and a “data event”, \( D=D \),

\[ P(H=h \mid D=D) = \frac{P(D=D \mid H=h) \cdot P(H=h)}{P(D=D)} \]

rewritten using probability mass functions, pmf, (in case of discrete events):

\[ p(h \mid D) = \frac{p(D \mid h) \cdot p(h)}{p(D)} \]

- Likelihood: How well does \( h \) explain (= entail, induce, evoke) the data \( D \)?
- Prior: How probable is the hypothesis \( h \) a priori (= in principle)?
- Normalization: How probable is the observation of the data \( D \)?
- Posterior: How probable is the hypothesis \( h \) when observing the data \( D \)?
When using the Bayes method for a predictor-response setting, then $p(D)$, $p(D) := P(D=D)$, is the probability of the data $D = x$. I.e., $D$ is a random vector whose domain is the feature space $X$.

When using the Bayes method for an outcome-only setting, then $p(D)$, $p(D) := P(D=D)$, is the probability of the data $D = \{y_1, \ldots, y_n\}$ or $D = \{c_1, \ldots, c_n\}$. I.e., $D$ is a random vector whose domain is $\mathbb{R}^n$ or $\mathbb{C}^n$, where $C$ is the set of possible classes or class labels.

$p(h) := P(H=h)$ (also $p(w)$, $p(\theta)$, or similar) is the probability of choosing a certain $h$, a parameter vector $w$, or some model function as hypothesis. I.e., $H$ is a random variable whose domain is the set $H$ of possible hypotheses.

Recall that $p()$ is defined via $P()$ and that the two notations can be used interchangeably, arguing about realizations of random variables and events respectively.