Chapter ML:II (continued)

II. Machine Learning Basics

- Regression
- Concept Learning: Search in Hypothesis Space
- Concept Learning: Search in Version Space
- Evaluating Effectiveness

Evaluating Effectiveness

True Misclassification Rate

Definition 8 (True Misclassification Rate)

Let X be a feature space with a finite number of elements. Moreover, let C be a set of classes, let $y : X \rightarrow C$ be a classifier, and let c be the target concept to be learned. Then the true misclassification rate, denoted as *Err*$^*(y)$, is defined as follows:

$$\textit{Err}^*(y) = \frac{|\{\mathbf{x} \in X : y(\mathbf{x}) \neq c(\mathbf{x})\}|}{|X|}$$

Evaluating Effectiveness

True Misclassification Rate

Definition 8 (True Misclassification Rate)

Let X be a feature space with a finite number of elements. Moreover, let C be a set of classes, let $y : X \rightarrow C$ be a classifier, and let c be the target concept to be learned. Then the true misclassification rate, denoted as *Err*$^*(y)$, is defined as follows:

$$\textit{Err}^*(y) = \frac{|\{\mathbf{x} \in X : y(\mathbf{x}) \neq c(\mathbf{x})\}|}{|X|}$$

Problem:

- Usually the *total function* c is unknown.

Solution:

- Estimation of *Err*$^*(y)$ with *Err*(y, D_{ts}), i.e., evaluation of y on a subset $D_{ts} \subseteq D$. Recall that for the feature vectors in D the target concept c is known.

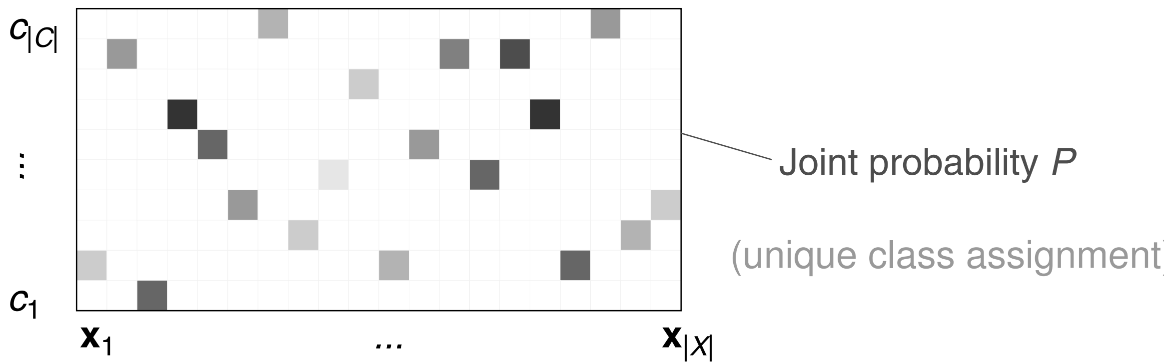
Remarks:

- Instead of the term “true misclassification rate” we may use also the term “true misclassification error” or simply “true error”.
- The English word “rate” can denote both the mathematical concept of a *flow quantity* (a change of a quantity per time unit) as well as the mathematical concept of a *portion*, a *percentage*, or a *ratio*, which has a stationary (= time-independent) semantics. Note that the latter semantics is meant here when talking about the misclassification rate.
- Unfortunately, the German word „Rate“ is often (mis)used to denote the mathematical concept of a portion, a percentage, or a ratio. Taking a precise mathematical standpoint, the correct German words are „Anteil“ or „Quote“. I.e., a semantically correct translation of misclassification rate is „Missklassifikationsanteil“, and not „Missklassifikationsrate“.

Evaluating Effectiveness

True Misclassification Rate: Probabilistic Foundation

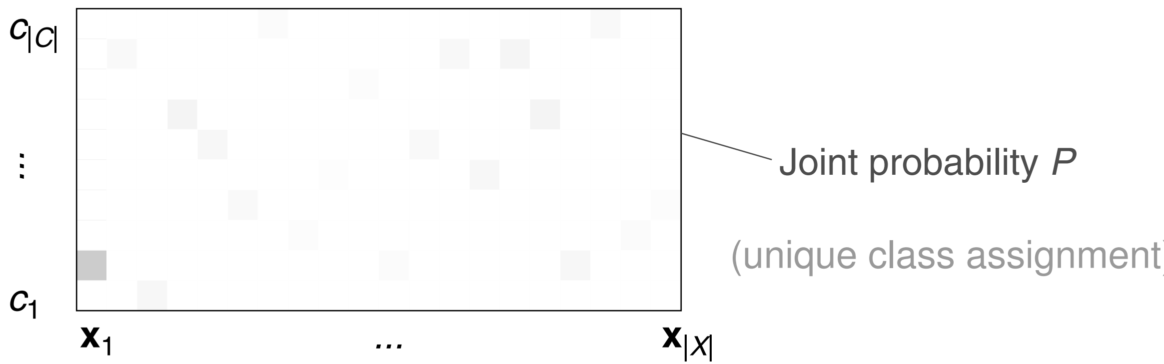
Let X be a feature space, C a set of classes, and P a probability measure on $X \times C$. Then $P(\mathbf{x}, c)$ (precisely: $P(\mathcal{X} = \mathbf{x}, \mathcal{C} = c)$) denotes the probability (1) to observe the vector $\mathbf{x} \in X$ and (2) that $\mathbf{x}$ belongs to class $c \in C$. Illustration:



Evaluating Effectiveness

True Misclassification Rate: Probabilistic Foundation (continued)

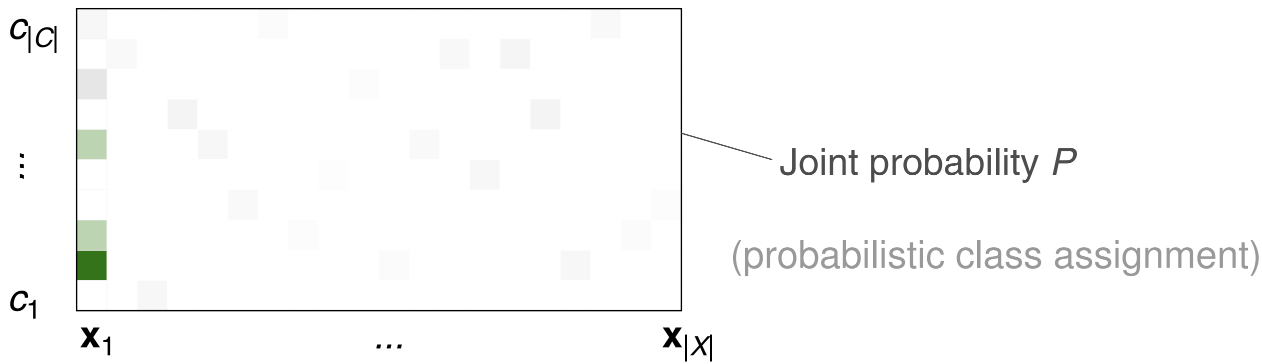
Let X be a feature space, C a set of classes, and P a probability measure on $X \times C$. Then $P(\mathbf{x}, c)$ (precisely: $P(\mathcal{X} = \mathbf{x}, \mathcal{C} = c)$) denotes the probability (1) to observe the vector $\mathbf{x} \in X$ and (2) that $\mathbf{x}$ belongs to class $c \in C$. Illustration:



Evaluating Effectiveness

True Misclassification Rate: Probabilistic Foundation (continued)

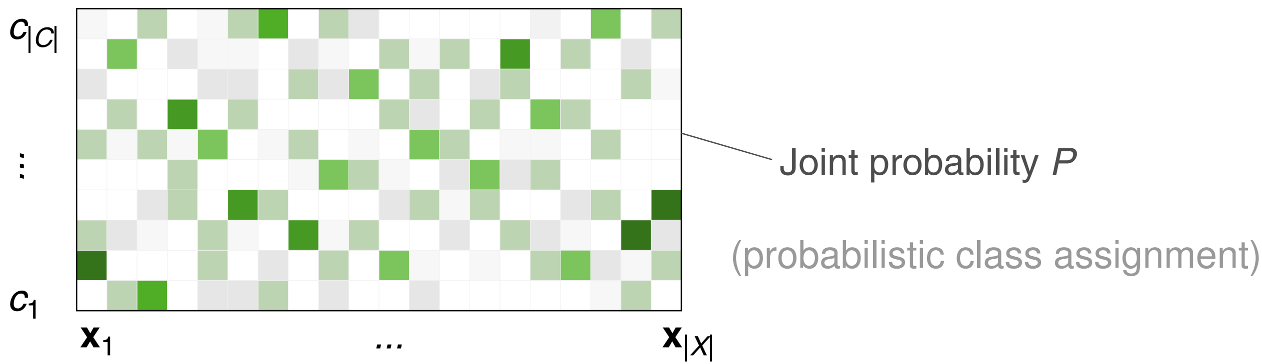
Let X be a feature space, C a set of classes, and P a probability measure on $X \times C$. Then $P(\mathbf{x}, c)$ (precisely: $P(\mathcal{X} = \mathbf{x}, \mathcal{C} = c)$) denotes the probability (1) to observe the vector $\mathbf{x} \in X$ and (2) that $\mathbf{x}$ belongs to class $c \in C$. Illustration:



Evaluating Effectiveness

True Misclassification Rate: Probabilistic Foundation (continued)

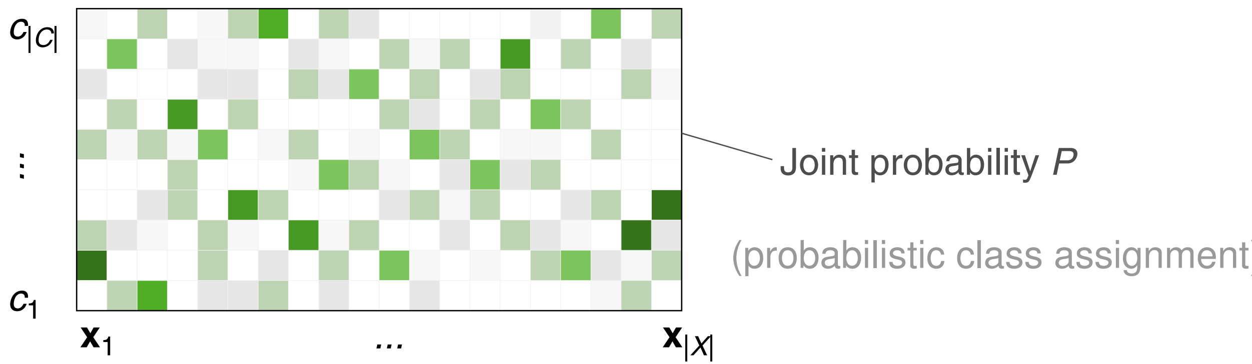
Let X be a feature space, C a set of classes, and P a probability measure on $X \times C$. Then $P(\mathbf{x}, c)$ (precisely: $P(\mathcal{X} = \mathbf{x}, \mathcal{C} = c)$) denotes the probability (1) to observe the vector $\mathbf{x} \in X$ and (2) that $\mathbf{x}$ belongs to class $c \in C$. Illustration:



Evaluating Effectiveness

True Misclassification Rate: Probabilistic Foundation (continued)

Let X be a feature space, C a set of classes, and P a probability measure on $X \times C$. Then $P(\mathbf{x}, c)$ (precisely: $P(\mathcal{X} = \mathbf{x}, \mathcal{C} = c)$) denotes the probability (1) to observe the vector $\mathbf{x} \in X$ and (2) that $\mathbf{x}$ belongs to class $c \in C$. Illustration:

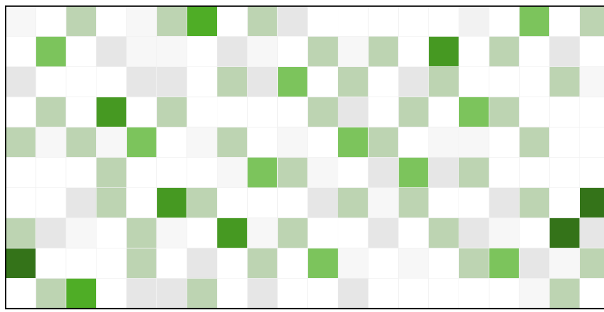


$$\mathit{Err}^*(y) = \sum_{\mathbf{x} \in X} \sum_{c \in C} P(\mathbf{x}, c) \cdot I(y(\mathbf{x}), c), \quad \text{with } I(y(\mathbf{x}), c) = \begin{cases} 0 & \text{if } y(\mathbf{x}) = c \\ 1 & \text{otherwise} \end{cases}$$

Evaluating Effectiveness

True Misclassification Rate: Probabilistic Foundation (continued)

Let X be a feature space, C a set of classes, and P a probability measure on $X \times C$. Then $P(\mathbf{x}, c)$ (precisely: $P(\mathcal{X} = \mathbf{x}, \mathcal{C} = c)$) denotes the probability (1) to observe the vector $\mathbf{x} \in X$ and (2) that $\mathbf{x}$ belongs to class $c \in C$. Illustration:



$$Err^*(y) = \sum_{\mathbf{x} \in X} \sum_{c \in C} P(\mathbf{x}, c) \cdot I(y(\mathbf{x}), c), \quad \text{with } I(y(\mathbf{x}), c) = \begin{cases} 0 & \text{if } y(\mathbf{x}) = c \\ 1 & \text{otherwise} \end{cases}$$

$D = \{(\mathbf{x}_1, c_1)), \ldots, (\mathbf{x}_n, c_n))\} \subseteq X \times C$, as well as $D_{ts} \subseteq D$, are sets of examples whose elements are drawn independently and according to P.

Remarks:

- $\mathcal{X}$ and $\mathcal{C}$ denote random variables with domains X and C respectively. In particular, X may not be restricted to a finite set.

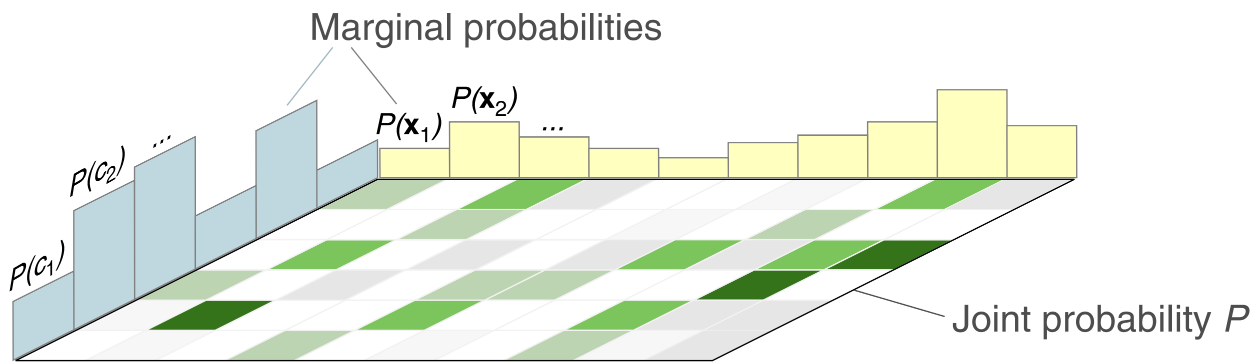
- $\mathcal{X}$ accounts for the fact that each observation process is governed by a probability distribution, rendering certain observations more likely than others.
  Note that in the definition of the True Misclassification Rate the elements in X are implicitly treated as uniformly distributed: each $\mathbf{x} \in X$ is considered with the same weight in *Err**.

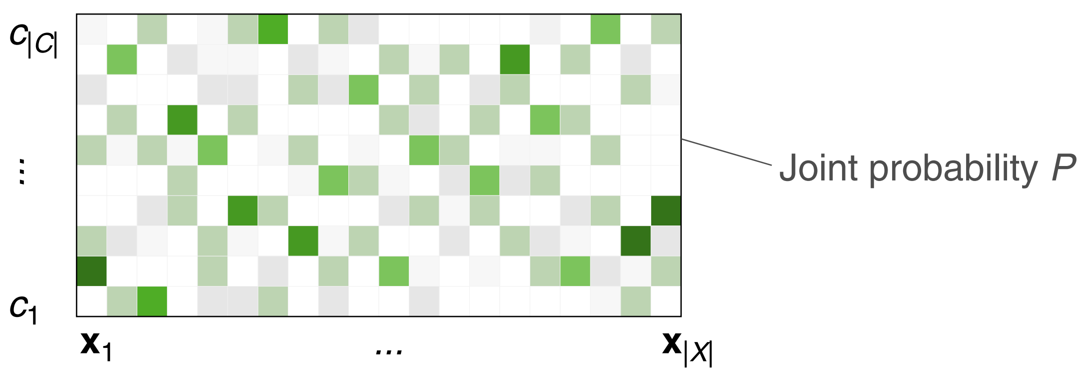
- $\mathcal{C}$ accounts for the fact that in the real world the classification $c(\mathbf{x})$ of a feature vector $\mathbf{x}$ may not be deterministic but the result of a random (measuring) process. Keyword: label noise

- If the elements in D and D_{ts} are not chosen according to P, then *Err*(y, D_{ts}) cannot be used as an estimation of *Err*$^*(y)$. Keyword: sample selection bias
  The fact that random variables are both independent of each other and identically distributed is abbreviated with "i.i.d."

- P is a probability measure and hence its argument must be an event, such as "$\mathcal{X} = \mathbf{x}$" or "$\mathcal{X} = \mathbf{x} \mid \mathcal{C} = c$". I.e., notations such as $P(\mathbf{x})$ and $P(\mathbf{x} \mid c)$ are abbreviations of $P(\mathcal{X} = \mathbf{x})$ and $P(\mathcal{X} = \mathbf{x} \mid \mathcal{C} = c)$ respectively.
  Let A and B denote two events, e.g., $A =$ "$\mathbf{x}$ [was observed]" and $B =$ "c [is the class of $\mathbf{x}$]".
  Then the following expressions are syntactic variants to denote the probability of the combined event: $P(A, B)$, $P(A$ and $B)$, $P(A \wedge B)$.

Illustration of the marginal probabilities $P(c)$ and $P(\mathbf{x})$:





Remarks:

- P_c (precisely: $P_{\mathcal{C}=c}$) is the probability distribution of the $\mathbf{x} \in X$ under class c.
  $P_c(\mathbf{x}) \;\equiv\; P(\mathcal{X} = \mathbf{x} \mid \mathcal{C} = c)$.
  P_c is a probability measure, also called "class-conditional *probability* *[density]* function".

  In the illustration: the distribution of the $\mathbf{x}$ (consider a row) for a certain class c.
  Summation / integration over the $\mathbf{x} \in X$ yields the marginal probability $P(c)$.

- $P_{\mathbf{x}}$ (precisely: $P_{\mathcal{X}=\mathbf{x}}$) is the probability distribution of the $c \in C$ under feature vector $\mathbf{x}$.
  $P_{\mathbf{x}}(c) \;\equiv\; P(\mathcal{C} = c \mid \mathcal{X} = \mathbf{x})$.
  $P_{\mathbf{x}}$ is a probability measure, also called "conditional *class probability* function".

  In the illustration: the distribution of the c (consider a column) for a certain feature vector $\mathbf{x}$.
  Summation over the $c \in C$ yields the marginal probability $P(\mathbf{x})$.

- $P(\mathbf{x}, c) \;=\; P(\mathbf{x} \mid c) \cdot P(c)$, where $P(c)$ is the a-priori probability for (observing) event c, and $P(\mathbf{x} \mid c)$ is the probability for (observing) event $\mathbf{x}$ given event c.

  Likewise, $P(\mathbf{x}, c) \;=\; P(c, \mathbf{x}) \;=\; P(c \mid \mathbf{x}) \cdot P(\mathbf{x})$, where $P(\mathbf{x})$ is the a-priori probability for (observing) event $\mathbf{x}$, and $P(c \mid \mathbf{x})$ is the probability for (observing) event c given event $\mathbf{x}$.

- Let both events $\mathbf{x}$ and c have occurred already, and, let $\mathbf{x}$ be known and c be unknown. Then, $P(\mathbf{x} \mid c)$ is called *likelihood* (for event $\mathbf{x}$ given event c).

Evaluating Effectiveness

Training Error [True Misclassification Rate]

- $D = \{(\mathbf{x}_1, c(\mathbf{x}_1)), \ldots, (\mathbf{x}_n, c(\mathbf{x}_n))\} \subseteq X \times C$ is a set of examples.
- $D_{tr} = D$ is the training set.
- $y : X \to C$ is a classifier learned on the basis of D_{tr}.

Training error = misclassification rate with respect to D_{tr} :

$$\mathit{Err}(y, D_{tr}) = \frac{|\{(\mathbf{x}, c(\mathbf{x})) \in D_{tr} : y(\mathbf{x}) \neq c(\mathbf{x})\}|}{|D_{tr}|}$$

Evaluating Effectiveness

Training Error [True Misclassification Rate]

- $D = \{(\mathbf{x}_1, c(\mathbf{x}_1)), \ldots, (\mathbf{x}_n, c(\mathbf{x}_n))\} \subseteq X \times C$ is a set of examples.
- $D_{tr} = D$ is the training set.
- $y : X \to C$ is a classifier learned on the basis of D_{tr}.

Training error = misclassification rate with respect to D_{tr} :

$$\mathit{Err}(y, D_{tr}) = \frac{|\{(\mathbf{x}, c(\mathbf{x})) \in D_{tr} : y(\mathbf{x}) \neq c(\mathbf{x})\}|}{|D_{tr}|}$$

Problems:

- $\mathit{Err}(y, D_{tr})$ is based on examples that are also exploited to learn y.

➜ $\mathit{Err}(y, D_{tr})$ quantifies memorization but not the generalization capability of y.

➜ $\mathit{Err}(y, D_{tr})$ is an optimistic estimation, i.e., it is constantly lower compared to the error incurred when applying y in the wild.

Evaluating Effectiveness

Holdout Estimation [True Misclassification Rate]

- $D = \{(\mathbf{x}_1, c(\mathbf{x}_1)), \ldots, (\mathbf{x}_n, c(\mathbf{x}_n))\} \subseteq X \times C$ is a set of examples.
- $D_{tr} \subset D$ is the training set.
- $y : X \to C$ is a classifier learned on the basis of D_{tr}.
- $D_{ts} \subset D$ with $D_{ts} \cap D_{tr} = \emptyset$ is a test set.

Holdout estimation = misclassification rate with respect to D_{ts} :

$$\mathit{Err}(y, D_{ts}) = \frac{|\{(\mathbf{x}, c(\mathbf{x})) \in D_{ts} : y(\mathbf{x}) \neq c(\mathbf{x})\}|}{|D_{ts}|}$$

Evaluating Effectiveness

Holdout Estimation [True Misclassification Rate]

- $D = \{(\mathbf{x}_1, c(\mathbf{x}_1)), \ldots, (\mathbf{x}_n, c(\mathbf{x}_n))\} \subseteq X \times C$ is a set of examples.
- $D_{tr} \subset D$ is the training set.
- $y : X \to C$ is a classifier learned on the basis of D_{tr}.
- $D_{ts} \subset D$ with $D_{ts} \cap D_{tr} = \emptyset$ is a test set.

Holdout estimation = misclassification rate with respect to D_{ts} :

$$\mathit{Err}(y, D_{ts}) = \frac{|\{(\mathbf{x}, c(\mathbf{x})) \in D_{ts} : y(\mathbf{x}) \neq c(\mathbf{x})\}|}{|D_{ts}|}$$

Requirements:

- D_{tr} and D_{ts} must be governed by the same distribution.
- D_{tr} and D_{ts} should have similar sizes.

Remarks:

- A typical value for splitting D into training set D_{tr} and test set D_{ts} is 2:1.
- When splitting D into D_{tr} and D_{ts} one has to ensure that the underlying distribution is maintained. Keywords: stratification, sample selection bias

Evaluating Effectiveness

k-Fold Cross-Validation [Holdout Estimation]

- Form k test sets by splitting D into disjoint sets $D_1, \ldots, D_k$ of similar size.
- For $i = 1, \ldots, k$ do:
  1. $y_i : X \to C$ is a classifier learned on the basis of $D \setminus D_i$
  2. $$\mathit{Err}(y_i, D_i) = \frac{|\{(\mathbf{x}, c(\mathbf{x}) \in D_i : y_i(\mathbf{x}) \neq c(\mathbf{x})\}|}{|D_i|}$$

Cross-validated misclassification rate:

$$\mathit{Err}_{cv}(y, D) = \frac{1}{k} \sum_{i=1}^{k} \mathit{Err}(y_i, D_i)$$

Evaluating Effectiveness

n-Fold Cross-Validation (Leave One Out)

Special case with $k = n$:

- Determine the cross-validated misclassification rate for $D \setminus D_i$ where $D_i = \{(\mathbf{x}_i, c(\mathbf{x}_i))\},\ i \in \{1, \ldots, n\}$.

Evaluating Effectiveness

n-Fold Cross-Validation (Leave One Out)

Special case with $k = n$:

- Determine the cross-validated misclassification rate for $D \setminus D_i$ where $D_i = \{(\mathbf{x}_i, c(\mathbf{x}_i))\},\ i \in \{1, \ldots, n\}$.

Problems:

- High computational effort if D is large.
- Singleton test sets ($|D_i| = 1$) are never stratified since they contain a single class only.

Remarks:

- For large k the set $D \setminus D_i$ is of similar size as D. Hence *Err*(y_i, D_i) is close to *Err*(y, D), where y is the classifier learned on the basis of the entire set D.
- n-fold cross-validation is a special case of exhaustive cross-validation methods, which learn and test on all possible ways to divide the original sample into a training and a validation set. [Wikipedia]

Evaluating Effectiveness

Misclassification Costs [Holdout Estimation]

Use of a cost measure for the misclassification of a feature vector $\mathbf{x}$ in class c' instead of in class c :

$$\textit{cost}(c' \mid c) \begin{cases} \geq 0 & \text{if } c' \neq c \\ = 0 & \text{otherwise} \end{cases}$$

Estimation of $\textit{Err}^*_{\textit{cost}}(y)$ based on a sample $D_{ts} \subseteq D$:

$$\textit{Err}_{\textit{cost}}(y, D_{ts}) = \frac{1}{|D_{ts}|} \cdot \sum_{(\mathbf{x}, c(\mathbf{x})) \in D_{ts}} \textit{cost}(y(\mathbf{x}) \mid c(\mathbf{x}))$$

Remarks:

- The misclassification rate *Err* is a special case of Err_{cost} with $cost(c' \mid c) = 1$ for $c' \neq c$.