Pattern Classification

Chapter 9.6 Estimating and Comparing Classifiers

Introduction I

- Two reasons to know the generalization rate of a classifier:
    - the classifier performs well enough to be useful.
    - to compare its performance with that of a competing design

Outline

1. Parametric models
2. Cross validation
3. Jackknife and bootstrap estimation of classification accuracy
4. Maximum-likelihood model comparison
5. Bayesian model comparison
6. The problem-average error rate
7. The capacity of a separating plane

Parametric model I

- One approach: To estimate the generalization rate from the assumed parametric model.
- 3 problems:
  - error estimate is often optimistic.
  - suspect the validity of an assumed parametric model.
  - it is very difficult to compute the error rate exactly, even if the probabilistic structure is known completely.

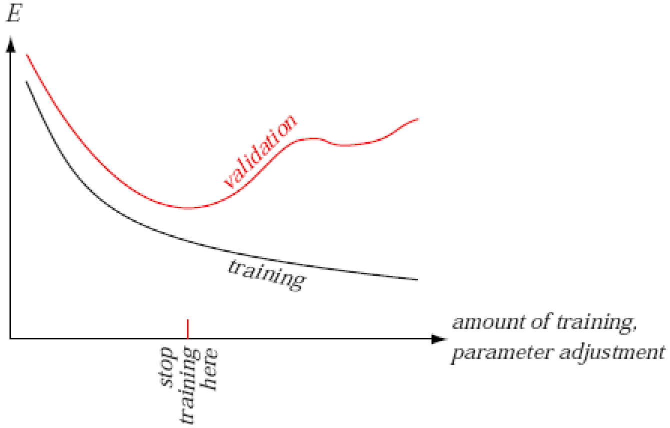
Outline

1. Parametric models
2. Cross validation
3. Jackknife and bootstrap estimation of classification accuracy
4. Maximum-likelihood model comparison
5. Bayesian model comparison
6. The problem-average error rate
7. The capacity of a separating plane

Cross validation I

- Randomly split the set of labeled training samples D into two parts:
    - Training set: for adjusting de parameters.
    - Validation set: estimate the generalization error.
- We train the classifier until set we reach a minimum of this validation error:

Cross validation II



Cross validation I

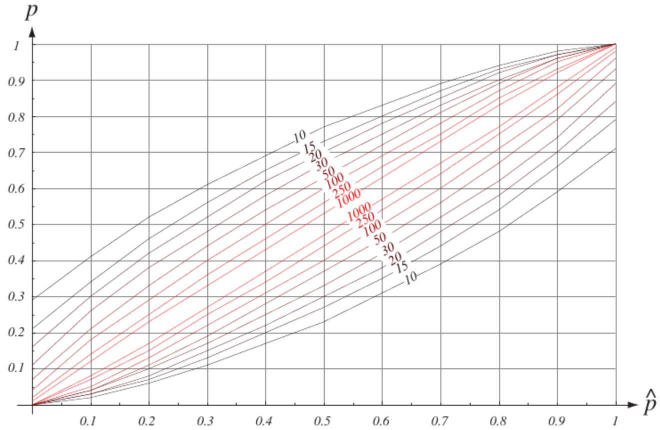
- Cross validation is heuristic and need not give improved classifiers in every case.
- There are several heuristics for choosing the portion γ of D to be used as a validation set ($0 < \gamma < 1$).
  - small portion of the data: validation set ($\gamma < 0.5$)
  - A traditional default is to split the data with $\gamma = 0.1$.
  - **m-fold cross validation**: the cross validation training set is randomly divided into ***m*** disjoint sets of equal size n/m. ($m{=}n$, leave-one-out)
  - **anti-cross validation:** stop training when the validation error is the first local maximum.
  - If the true but unknown error rate of the classifier is p, and if k of the n independent, randomly drawn test samples are misclassified, then k has the binomial distribution

Cross validation II

$$P(k) = \binom{n'}{k} p^k (1-p)^{n'-k}. \qquad \hat{p} = \frac{k}{n'}.$$

the fraction of test samples misclassified is exactly the maximum likelihood estimate for *p*.

Cross validation I



- The 95% confidence intervals for a given estimated error probability $\hat{p}$ can be derived from a binomial distribution of equation P(k).

Outline

1. Parametric models
2. Cross validation
3. Jackknife and bootstrap estimation of classification accuracy
4. Maximum-likelihood model comparison
5. Bayesian model comparison
6. The problem-average error rate
7. The capacity of a separating plane

Jackknife and bootstrap estimation of classification accuracy I

- **Jackknife:** we estimate the accuracy of a given algorithm by training the classifier n separate times, each time using the training set D from which a different single training point has been deleted. Each resulting classifier is tested on the single deleted point and the jackknife estimate of the accuracy is then simply the mean of these leave-one-out accuracies.
- There are several ways to generalize the bootstrap method to the problem of estimating the accuracy of a classifier. One of the simplest approaches is to train B classifiers, each with a different bootstrap data set, and test on other bootstrap data sets.
- The bootstrap estimate of the classifier accuracy is simply the mean of these bootstrap accuracies.

Outline

1. Parametric models
2. Cross validation
3. Jackknife and bootstrap estimation of classification accuracy
4. Maximum-likelihood model comparison
5. Bayesian model comparison
6. The problem-average error rate
7. The capacity of a separating plane

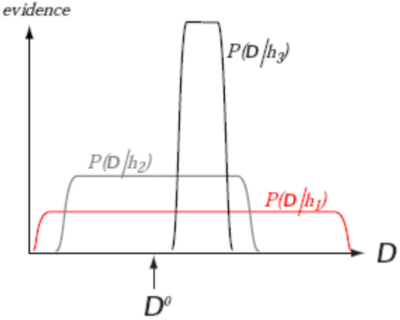
Maximum-likelihood model comparison I

- Maximum-likelihood model comparison (ML-II): Given a model with unknown parameter vector θ, we find the value $\hat{\theta}$ which maximizes the probability of the training data. The goal here is to choose the model that best explains the training data
- The posterior probability of any given model:

$$P(h_i|\mathcal{D}) = \frac{P(\mathcal{D}|h_i)P(h_i)}{p(\mathcal{D})} \propto P(\mathcal{D}|h_i)P(h_i),$$

- The data-dependent term, $P(D|h_i)$, is the evidence for h_i; the second term, $P(h_i)$, is our subjective prior over the space of hypotheses.

Maximum-likelihood model comparison II



Outline

1. Parametric models
2. Cross validation
3. Jackknife and bootstrap estimation of classification accuracy
4. Maximum-likelihood model comparison
5. Bayesian model comparison
6. The problem-average error rate
7. The capacity of a separating plane

Bayesian model comparison I

- Uses the full information over priors when computing posterior probabilities.
- The evidence for a particular hypothesis is an integral,

$$P(\mathcal{D}|h_i) = \int p(\mathcal{D}|\boldsymbol{\theta}, h_i)p(\boldsymbol{\theta}|\mathcal{D}, h_i)d\boldsymbol{\theta}, \tag{41}$$

where as before θ describes the parameters in the candidate model.

$$P(\mathcal{D}|h_i) \simeq \underbrace{P(\mathcal{D}|\hat{\boldsymbol{\theta}}, h_i)}_{\substack{\text{best fit}\\ \text{likelihood}}} \; \underbrace{p(\hat{\boldsymbol{\theta}}|h_i)\Delta\boldsymbol{\theta}}_{\text{Occam factor}} \quad .$$

Bayesian model comparison I

$$
\begin{aligned}
\text{Occam factor} \quad &= \quad p(\hat{\theta}|h_i)\Delta\theta = \frac{\Delta\theta}{\Delta^0\theta} \\
&= \quad \frac{\text{param. vol. commensurate with } \mathcal{D}}{\text{param. vol. commensurate with any data}},
\end{aligned}
$$

is the ratio of two volumes in parameter space:

1. the volume that can account for data D and
2. the prior volume, accessible to the model without regard to D.

Bayesian model comparison I

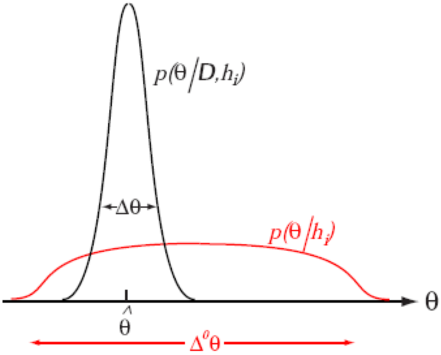


Figure 9.13: In the absence of training data, a particular model h has available a large range of possible values of its parameters, denoted $\Delta^0\theta$. In the presence of a particular training set $\mathcal{D}$, a smaller range is available. The Occam factor, $\Delta\theta/\Delta^0\theta$, measures the fractional decrease in the volume of the model's parameter space due to the presence of training data $\mathcal{D}$. In practice, the Occam factor can be calculated fairly easily if the evidence is approximated as a k-dimensional Gaussian, centered on the maximum-likelihood value $\hat{\boldsymbol{\theta}}$.

Bayesian model comparison I

In the general case, the full integral of Eq. 41 is too difficult to calculate analytically or even numerically. Nevertheless, if $\boldsymbol{\theta}$ is k-dimensional and the posterior can be assumed to be a Gaussian, then the Occam factor can be calculated directly (Problem 37), yielding:

$$P(\mathcal{D}|h_i) \simeq \underbrace{P(\mathcal{D}|\hat{\boldsymbol{\theta}}, h_i)}_{\substack{\text{best fit}\\ \text{likelihood}}} \underbrace{p(\hat{\boldsymbol{\theta}}|h_i)(2\pi)^{k/2}|\mathbf{H}|^{-1/2}}_{\text{Occam factor}}. \qquad (44)$$

where

$$\mathbf{H} = \frac{\partial^2 \ln p(\boldsymbol{\theta}|\mathcal{D}, h_i)}{\partial \boldsymbol{\theta}^2} \qquad (45)$$

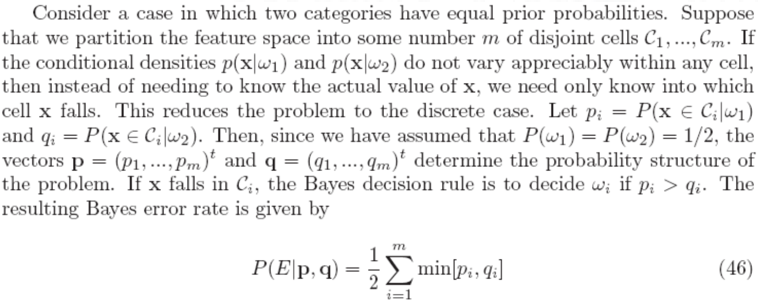
Outline

1. Parametric models
2. Cross validation
3. Jackknife and bootstrap estimation of classification accuracy
4. Maximum-likelihood model comparison
5. Bayesian model comparison
6. The problem-average error rate
7. The capacity of a separating plane

The problem-average error rate I

- Having only a small number of samples is that the resulting classifier will not perform well on new data.
- We expect the error rate to be a function of the number n of training samples
- To investigate this analytically:
    - Estimate the unknown parameters from samples.
    - Use these estimates to determine the classifier.
    - Calculate the error rate for the resulting classifier.

The problem-average error rate I



Consider a case in which two categories have equal prior probabilities. Suppose that we partition the feature space into some number m of disjoint cells $\mathcal{C}_1, ..., \mathcal{C}_m$. If the conditional densities $p(\mathbf{x}|\omega_1)$ and $p(\mathbf{x}|\omega_2)$ do not vary appreciably within any cell, then instead of needing to know the actual value of $\mathbf{x}$, we need only know into which cell $\mathbf{x}$ falls. This reduces the problem to the discrete case. Let $p_i = P(\mathbf{x} \in \mathcal{C}_i|\omega_1)$ and $q_i = P(\mathbf{x} \in \mathcal{C}_i|\omega_2)$. Then, since we have assumed that $P(\omega_1) = P(\omega_2) = 1/2$, the vectors $\mathbf{p} = (p_1, ..., p_m)^t$ and $\mathbf{q} = (q_1, ..., q_m)^t$ determine the probability structure of the problem. If $\mathbf{x}$ falls in $\mathcal{C}_i$, the Bayes decision rule is to decide ω_i if $p_i > q_i$. The resulting Bayes error rate is given by

$$P(E|\mathbf{p}, \mathbf{q}) = \frac{1}{2}\sum_{i=1}^{m} \min[p_i, q_i] \qquad (46)$$

The problem-average error rate I

- Suppose that half of the samples are labeled ω_1 and half are labeled ω_2 , with n_{ij} being the number that fall in C_i and are labeled ω_j .
- $\hat{p}_i = 2n_{i1}/n$ and $\hat{q}_i = 2n_{i2}/n$

$$P(E|\mathbf{p},\mathbf{q},\mathcal{D}) = \frac{1}{2}\sum_{n_{i1}>n_{i2}} q_i + \frac{1}{2}\sum_{n_{i1}\leq n_{i2}} p_i.$$

The problem-average error rate I

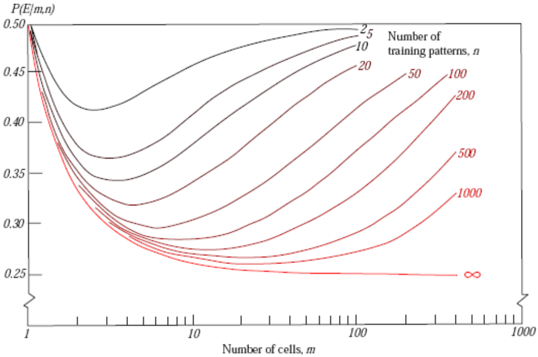
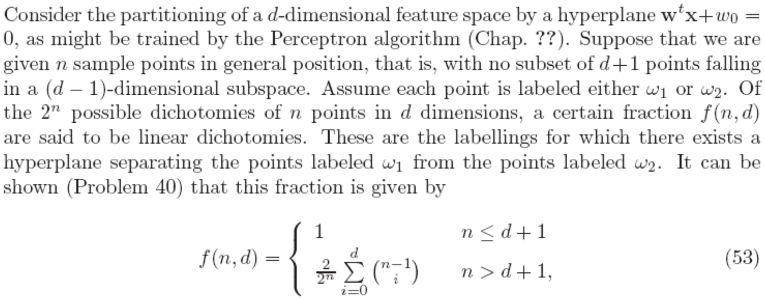


Figure 9.14: The probability of error E on a two-category problem for a given number of samples, n, can be estimated by splitting the feature space into m cells of equal size and classifying a test point by according to the label of the most frequently represented category in the cell. The graphs show the average error of a large number of random problems having the given n and m indicated.

Outline

1. Parametric models
2. Cross validation
3. Jackknife and bootstrap estimation of classification accuracy
4. Maximum-likelihood model comparison
5. Bayesian model comparison
6. The problem-average error rate
7. The capacity of a separating plane

The capacity of a separating plane I



Consider the partitioning of a d-dimensional feature space by a hyperplane $\mathbf{w}^t\mathbf{x}+w_0 = 0$, as might be trained by the Perceptron algorithm (Chap. ??). Suppose that we are given n sample points in general position, that is, with no subset of $d+1$ points falling in a $(d-1)$-dimensional subspace. Assume each point is labeled either ω_1 or ω_2. Of the 2^n possible dichotomies of n points in d dimensions, a certain fraction $f(n,d)$ are said to be linear dichotomies. These are the labellings for which there exists a hyperplane separating the points labeled ω_1 from the points labeled ω_2. It can be shown (Problem 40) that this fraction is given by

$$f(n,d) = \begin{cases} 1 & n \leq d+1 \\ \frac{2}{2^n} \sum\limits_{i=0}^{d} \binom{n-1}{i} & n > d+1, \end{cases} \tag{53}$$

The capacity of a separating plane

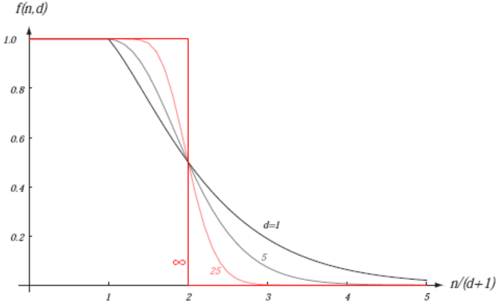


Figure 9.18: The fraction of dichotomies of n points in d dimensions that are linear, as given by Eq. 53.