Pattern Classification

Chapter 9.6 Estimating and Comparing Classifiers

- 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

Parametric model I

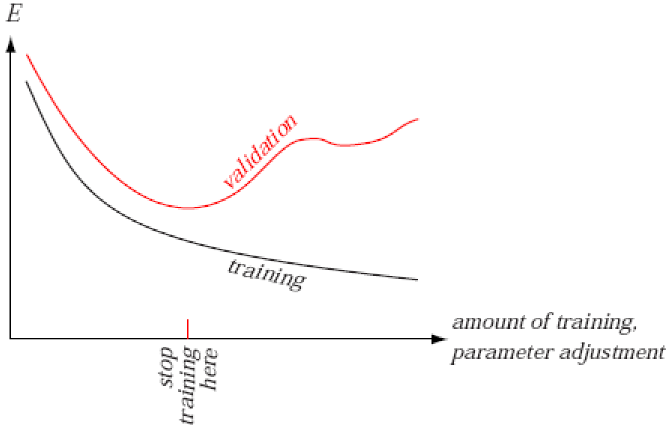
- One approach: To estimate the generalization rate from the assumed parametric model.
- 3 problems:
  - error estimate is often optimistic.
  - suspect the validity of an assumed parametric model.
  - it is very difficult to compute the error rate exactly, even if the probabilistic structure is known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

Outline

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

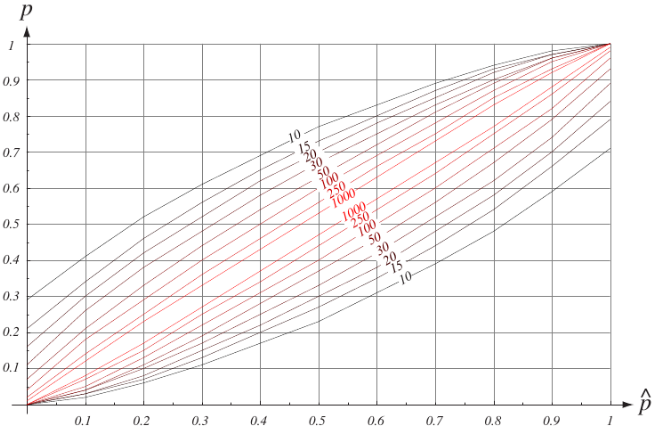
- Cross validation is heuristic and need not give improved classifiers in every case.
- There are several heuristics for choosing the portion γ of D to be used as a validation set $(0 < \gamma < 1)$.
  - small portion of the data: validation set $(\gamma < 0.5)$
  - A traditional default is to split the data with $\gamma = 0.1$.
  - **m-fold cross validation**: the cross validation training set is randomly divided into $\boldsymbol{m}$ 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

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

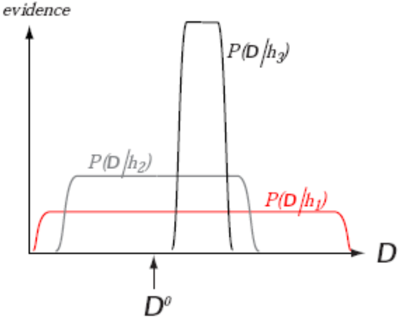
Maximum-likelihood model comparison I

- 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

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.

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

Bayesian model comparison I

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

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

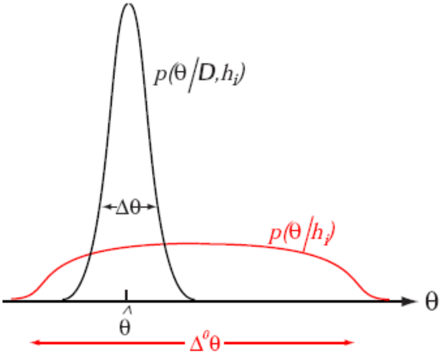


Figure 9.13: In the absence of training data, a particular model h has available a large range of possible values of its parameters, denoted $\Delta^0\theta$. In the presence of a particular training set $\mathcal{D}$, a smaller range is available. The Occam factor, $\Delta\theta/\Delta^0\theta$, measures 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 $\hat{\theta}$.

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:

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

where

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