EPFL, Autumn 2024
Yoav Zemel
Guillaume Obozinski

Ho Yun
Shivang Sachar

Statistical Machine Learning

Exercise sheet 5

Exercise 5.1 (Leave-one-out cross-validation for *linear smoothers*) In this exercise we consider *linear smoothers*, i.e., learning scheme producing decision functions $\widehat{f}$ for which the fitted values $\widehat{y}_i := \widehat{f}(\boldsymbol{x}_i)$ on the training set satisfy $\widehat{\boldsymbol{y}} = \mathbf{S}\boldsymbol{y}$, where $\mathbf{S}$ is an $n \times n$ matrix whose values only depend on the inputs $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ and $\widehat{\boldsymbol{y}} = (y_i)_{i=1\dots n}$.

We consider the leave-one-out CV error

$$\mathrm{CV}(\widehat{f}) = \frac{1}{n}\sum_{i=1}^{n}\left\{y_i - \widehat{f}^{-i}(\boldsymbol{x}_i)\right\}^2,$$

where $\widehat{f}^{-i}$ denote the model fitted to the original training sample with the ith observation $(y_i, \boldsymbol{x}_i)$ removed.

The goal of this exercise is to derive a fast way of computing the leave-one-out (or n-fold) cross-validation (CV) error for *linear smoothers* which produce leave-one-out decision functions with a particular form (given by Equation (1) below).

(a) Show that linear regression is a linear smoother in the sense that the obtained prediction function $\widehat{f}$ satisfies the property above. In particular specify $\mathbf{S}$.

(b) Assume that the leave-ith-out fit at $\boldsymbol{x}_i$ is given by

$$\widehat{f}^{-i}(\boldsymbol{x}_i) = \sum_{j\neq i}\frac{\mathbf{S}_{ij}}{1-\mathbf{S}_{ii}}y_j. \tag{1}$$

With this regularity assumption, show that

$$y_i - \widehat{f}^{-i}(\boldsymbol{x}_i) = \frac{y_i - \widehat{f}(\boldsymbol{x}_i)}{1-\mathbf{S}_{ii}}. \tag{2}$$

(c) Explain why (2) may be used to compute the CV error more efficiently.

(d) Our goal in the rest of this exercise is to identify some conditions that imply that $\widehat{f}^{-i}$ is of the form (1). We consider the squared loss $\ell(a, y) = (a - y)^2$ and we focus on the decision function minimizing the empirical risk in a hypothesis class S, that is

$$\widehat{f} = \arg\min_{f\in S} \tfrac{1}{n}\sum_{i=1}^{n}(f(\boldsymbol{x}_i) - y_i)^2,$$

assuming that the latter is unique. Assume that $\widehat{f}^{-i}$ has been computed and that we define a new dataset $\tilde{D}_n = \{(\boldsymbol{x}_j, \tilde{y}_j)\}_{j=1\dots n}$ with $\tilde{y}_j = y_j$ for all $j \neq i$ and $\tilde{y}_i = \widehat{f}^{-i}(\boldsymbol{x}_i)$. Show that the minimizer of the empirical risk on this new dataset is $\widehat{f}^{-i}$.

(e) Given that the linear regression estimator is a linear smoother, there is a matrix $\mathbf{S}$ such that $\widehat{\boldsymbol{y}} = \mathbf{S}\boldsymbol{y}$. Use the previous question to show that $(\mathbf{S}\tilde{\boldsymbol{y}})_i = \widehat{f}^{-i}(\boldsymbol{x}_i)$ and use the form of $\tilde{\boldsymbol{y}}$ to prove that $\widehat{f}^{-i}$ takes the form of (1).

(f) Deduce from the previous questions the form of the LOO CV error for linear regression.

(g) Can a similar approach be used to obtain an expression of the LOO CV error for ridge regression?

(h) Show that all local averaging methods are linear smoothers.

(i) Show that (1) holds for the Nadaraya-Watson estimator, and deduce the LOO CV error for it.

(j) Does (1) hold for histogram estimators? For the k nearest-neighbors?

Exercise 5.2 (Fisher Discriminant) Logistic regression was introduced in class as an optimization problem which is obtained by applying the maximum likelihood principle to a model of $p(y = 1|x)$ in which the log-odd ratio is an affine function of the input feature vector. This type of model is often called *conditional model* or *discriminative* model because it only models the conditional distribution of y given x and not the marginal distribution of x. By contrast, we consider here what is called a *generative* model, a model in which both a model of $p(y)$ and $p(x|y)$ are estimated and from which $p(y|x)$ can be deduced (and also $p(x)$ of course). The particular models that we will consider are due to Fisher and are called *linear discriminant analysis* (LDA) and *quadratic discriminant analysis* (QDA). We will focus on the binary classification setting, although the method generalizes immediately to the multiclass classification setting.

(a) We first consider the QDA model. Given the class variable $y \in \{0, 1\}$, the data are assumed to be Gaussian with different means and different covariance matrices for the two different classes but with the same covariance matrix.

$$y \sim \text{Bernoulli}(\pi), \quad x|\{y = k\} \sim \text{Normal}(\mu_k, \Sigma_k),$$

with $x, \mu_k \in \mathbb{R}^p$ and $\Sigma_k \in \mathbb{R}^{p\times p}$. Derive the form of the maximum likelihood estimators for the parameters in this model, i.e. for $\pi, \mu_1, \mu_0, \Sigma_1$ and Σ_0.

(b) Give an expression of the conditional distribution $p(y = 1|x)$ as a function of $\pi, \mu_1, \mu_2, \Sigma_1$ and Σ_2.

(c) What is the equation of the classification boundary, i.e., of the set of points for which $p(y = 1|x) = 0.5$?

(d) LDA model. Given the class variable $y \in \{0, 1\}$, the data is now assumed to be Gaussian with different means for different classes but with the same covariance matrix.

$$y \sim \text{Bernoulli}(\pi), \quad x|\{y = i\} \sim \text{Normal}(\mu_k, \Sigma)$$

What is the maximum likelihood estimator for Σ now?

.

(e) What is the equation of the classification boundary, i.e., of the set of points for which $p(y = 1|x) = 0.5$? Compare the obtained predictor with the form of the logistic regression predictor.