Wissenschaftliches Rechnen II/Scientific Computing II

Sommersemester 2016
Prof. Dr. Jochen Garcke
Dipl.-Math. Sebastian Mayer

universitätbonn

Exercise sheet 7

To be handed in on **Thursday, 07.06.2016**

1 Model Selection

G 1. (Bias-variance decomposition)

Consider the squared loss $\ell_2(y,t) = (y-t)^2$. Let $D = \{(X_1,Y_1),\dots,(X_n,Y_n)\}$. The *expected test error for a fixed new input* $X = x$ of a learning method L is given by

$$\operatorname{err}(L,x) := \mathbb{E}[\ell_2(Y,\hat{f}(X))|X=x] = \mathbb{E}[\ell_2(Y,\hat{f}(x))],$$

where $\hat{f}(X) = L(D)(X)$. Show that $\operatorname{err}(L,x)$ can be decomposed as follows

$$\operatorname{err}(L,x) = \sigma^2 + (\operatorname{bias}(L,x))^2 + \operatorname{var}(L,x)$$

with *irreducible error* σ, *bias* term $\operatorname{bias}(L,x) = f(x) - \mathbb{E}[\hat{f}(x)]$, and *variance* term $\operatorname{var}(L,x) = \mathbb{E}[(\hat{f}(x) - \mathbb{E}[\hat{f}(x)])^2]$. Try to explain what kind of error each of the three terms describe.

Solution. To obtain the decomposition, use that ε is independent of the random training data D and add and subtract $\mathbb{E}[\hat{f}(x)]$:

$$\mathbb{E}[(f(x)+\varepsilon-\hat{f}(x)^2] = \mathbb{E}[\varepsilon^2] + \mathbb{E}[(f(x) - \mathbb{E}[\hat{f}(x)] + \mathbb{E}[\hat{f}(x)] - \hat{f}(x))^2].$$

The rest is simply calculating. The irreducible error cannot be avoided due to the noise in new observations. The bias is the approximation error which the learning method makes on average (so you train over and over again with different training data, average all obtained fits and consider the difference to $f(x)$). The variance provides a measure for how much the various fits obtained for different training data deviate from the average fit. □

G 2. (Extra- vs. in-sample error)

Let $\tilde{Y}_1,\dots,\tilde{Y}_n$ be an independent copy of $Y_1,\dots,Y_n$ and L some learning method. Let

$$R_{\ell_2,\text{in}}(f) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\ell_2(\tilde{Y}_i, f(x_i))]$$

be the *in-sample risk*. The *expected in-sample error* for given sampling points $x_1,\dots,x_n$ is defined as

$$\operatorname{err}_{\text{in}}(L,x_1,\dots,x_n) = \mathbb{E}[R_{\ell_2,\text{in}}(L(D)) \mid X_1 = x_1,\dots,X_n = x_n],$$

where D is defined in G1.

a) Let $P = P_{Y|X}\cdot P_X$ and $\hat{f} = L(D_{\text{train}})$. What is the difference between the risk $R_{\ell_2,P}(\hat{f})$, the empirical risk $R_{\ell_2,\text{emp}}(\hat{f})$, the in-sample risk $R_{\ell_2,\text{in}}(\hat{f})$, the expected in-sample error $\operatorname{err}_{\text{in}}(L,x,\dots,x_n)$, and the expected test error $\operatorname{err}(L,x)$.

b) Show that $\mathbb{E}[R_{\ell_2,P}(L(D))] = \mathbb{E}[\mathrm{err}(L, X)]$.

c) Let $\hat{f} = L(D)$. Show that

$$\mathrm{err}_{\mathrm{in}}(L, x_1, \ldots, x_n) = \mathbb{E}[R_{\ell_2,\mathrm{emp}}(\hat{f}) \mid X_1 = x_1, \ldots, X_n = x_n] + \frac{2}{N}\sum_{i=1}^{N} \mathrm{cov}(Y_i, \hat{f}(x_i)).$$

Solution.

a) The risk $R_{\ell_2,P}(\hat{f})$ is mean prediction error over all potential future observations $(x_{\mathrm{new}}, y_{\mathrm{new}})$ (represented by the random tuple (X, Y)) when we only consider the given training data D_{train}. The empirical risk is the average prediction error the fit $\hat{f}$ makes in reproducing the training observations $y_1, \ldots, y_n$. The in-sample risk is the average prediction error the fit $\hat{f}$ (for fixed training data D_{train}) makes averaged over all potential new sample values $\tilde{y}_1, \ldots, \tilde{y}_n$ (represented by taking the expectation of the random variables $\tilde{Y}_1, \ldots, \tilde{Y}_n$) at the given sample points $x_1, \ldots, x_n$. The expected in-sample error $\mathrm{err}_{\mathrm{in}}(L, x, \ldots, x_n)$ is the mean in-sample risk over all potential training samples values (represented by the random variables $Y_1, \ldots, Y_n$) while keeping the sample points $x_1, \ldots, x_n$ fixed.

b) Simply use that (X, Y) is independent of D, which allows you to interchange the order of taking expectations.

c) Simple calculating gives

$$\mathrm{err}_{\mathrm{in}}(L, x_1, \ldots, x_n) - \mathbb{E}[R_{\ell_2,\mathrm{emp}}(\hat{f}) \mid X_1 = x_1, \ldots, X_n = x_n] = \frac{2}{n}\sum_{i=1}^{n} \mathbb{E}[(Y_i - \mathbb{E}[Y_i])\hat{f}(x_i)].$$

Now we add add and substract $\mathbb{E}[\hat{f}(x_i)]$ to obtain

$$\begin{aligned}
&\mathbb{E}[(Y_i - \mathbb{E}[Y_i])\hat{(}f(x_i) - \mathbb{E}[\hat{f}(x_i)] + \mathbb{E}[\hat{f}(x_i)]] \\
&= \mathbb{E}[(Y_i - \mathbb{E}[Y_i])(\hat{f}(x_i) - \mathbb{E}[\hat{f}(x_i)]) + \mathbb{E}[(Y_i - \mathbb{E}[Y_i])\mathbb{E}[\hat{f}(x_i)]] \\
&= \mathrm{cov}(Y_i, \hat{f}(x_i)).
\end{aligned}$$

□

2 Support Vector Machines

G 3. (Geometrical interpretation of slack variables)

Consider the *soft margin SVM*

$$\min_{w\in\mathbb{R}^2, b\in\mathbb{R}} \frac{1}{2}w^T w + C\sum_{j=1}^{n}\xi_j \quad s.t. \quad y_i(w^T x^i + b) \geq 1 - \xi_i, \qquad i = 1, \ldots, n,$$
$$\xi_i \geq 0, \qquad i = 1, \ldots, n.$$

Give a geometric interpretation of the slack variables $\xi_1, \ldots, \xi_n$. To this end, fix some feasible vector $w \in \mathbb{R}^2$, assuming w.l.o.g. $w_1 < 0$ and $w_2 > 0$. Furthermore, consider w.l.o.g. the first data points $(x, y) = (x^1, y^1)$ with $x = (x_1, x_2)$. Now derive the geometrical interpretation by considering when $\xi > 0$ is required in the linear constraint $y(w^T x + b) \geq 1 - \xi$.

Solution. If $y = 1$ then $\xi > 0$ if and only if

$$x_2 < \frac{1-b}{w_2} - \frac{w_1}{w_2} x_1.$$

In an analogous manner, in case that $y = 1$ then $\xi > 0$ iff

$$x_2 > \frac{-1-b}{w_2} - \frac{w_1}{w_2} x_1.$$

Assume now that (x, y) is a data point correctly classified by w. Then the slack variable is nonzero if x lies inside a tube of width $2/w_2$ in x_2-direction around the seperating hyperplane given by $x_2 = -\frac{b}{w_2} - \frac{w_1}{w_2} x_1$. Consequently, the objective function $\min_{w \in \mathbb{R}^2} w^T w + C\xi$ penalizes *margin errors* (i.e., points inside the tube or lying in the wrong affine hyperplane) and not only *classification errors* (i.e., points lying in the wrong affine hyperplane). □