# 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

In previous programming exercises you dealt with the regularization parameter, k-fold cross valdiation to choose a good regularization parameter, and the computation of error measures. This usage of cross validation and the computation of errors was somewhat adhoc. We will use this exercise sheet and its programming exercises to put the estimation of the regularization parameter on more solid grounds from the viewpoint of *statistical learning theory*. To this end, imagine the following situation. You are given some data $D_{\text{train}} = \{(x_1, y_1), \dots, (x_1, y_n)\}$, $x_i \in \Omega \subseteq \mathbb{R}$, $y_i \in \mathbb{R}$, and you conjecture that there is some function $f : \Omega \to \mathbb{R}$ which explains the data, i.e., the sample value $y_i$ is the function value $f(x_i)$ perturbed by noise. Now you want to use some regularized kernel regression procedure to learn $f$. There are two different problems you have to address:

- **Model selection:** estimating the performance of different regularization parameters in order to choose the best one.
- **Model assessment:** having chosen a final model, estimating its prediction error (generalization error) on new data.

In order to do model selection and assessment, we use the following statistical model how the data has been generated. We assume that the data $(x_1, y_1), \dots, (x_n, y_n)$ are realizations of $n$ i.i.d. copies $(X_1, Y_1), \dots, (X_n, Y_n)$ of $(X, Y)$, where $X$ is a random variable taking values in $\Omega$ and $Y$ is a random variable taking values in $\mathbb{R}$, which is given by

$$Y = f(X) + \varepsilon,$$

where $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, $\sigma > 0$ fixed, is independent of $X$. The regularized kernel regression procedure can now be considered as a *learning method* $L : \bigcup_{n \in \mathbb{N}} (\Omega \times \mathbb{R})^n \to \{f : \Omega \to \mathbb{R}\}$, which maps given training data $D_{\text{train}}$ to a regression fit $L(D_{\text{train}}) = \hat{f} : \Omega \to \mathbb{R}$.

**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

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

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

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

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

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

Let $\tilde{Y}_1, \ldots, \tilde{Y}_n$ be an independent copy of $Y_1, \ldots, Y_n$ 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 $x_1, \ldots, x_n$ is defined as

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

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 $\text{err}_{\text{in}}(L, x, \ldots, x_n)$, and the expected test error $\text{err}(L, x)$.

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

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

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

## 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 $y(w^T x + b) \geq 1 - \xi$.

## 3 Homework

**H 1.** Let $\mathcal{H}$ be a Hilbertspace over $\Omega \subseteq \mathbb{R}$ and $k : \Omega \times \Omega \to \mathbb{R}$ the reproducing kernel of $\mathcal{H}$. Consider once again the regularized kernel regression problem

$$L(D_{\text{train}}) = \text{argmin}_{f\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda\|f\|_k^2$$

for given data $D_{\text{train}} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$.

a) Give explicit formulas for the bias and the variance term from G1.

b) Let $K = (k(x_i, x_i))_{i,j=1,\ldots,n}$ and $G = (K + \lambda I_n)^{-1}$. Consider the smoothing matrix $S_\lambda = KG$. Show that for the covariance term $\sum_{i=1}^{N} \text{cov}(Y_i, \hat{f}(x_i))$, which appears in G2 c), we have

$$\sum_{i=1}^{N} \text{cov}(Y_i, \hat{f}(x_i)) = \text{trace}(S_\lambda)\sigma.$$

(5 Punkte)

**H 2.** ($\nu$-support vector classifer)

For $\nu > 0$, consider the primal problem for the *$\nu$-SV classifier*

$$\min_{w\in\mathcal{H},\xi\in\mathbb{R}^n,\rho\in\mathbb{R}} \frac{1}{2}\|w\|_k^2 - \nu\rho + \frac{1}{n}\sum_{j=1}^{n}\xi_j \quad s.t. \quad y_i(\langle w, x_i\rangle) \geq \rho - \xi$$
$$\xi_i \geq 0, \rho \geq 0.$$

Note that $x_i$ denotes the feature representation of the $i$th sample point. The empirical *margin error* for a feasible $w \in \mathcal{H}$ is given by

$$R^\rho(w) := \frac{1}{n}|\{i \in \{1,\ldots,n\} : y_i\langle w, x_i\rangle < \rho\}|.$$

a) Suppose the above minimization problem has a solution with $\rho > 0$. Show that $\nu$ is an upper bound on the fraction of margin errors and a lower bound on the fraction of support vectors.

b) Determine the dual quadratic optimization problem. **Hint:** Use that the solution has a representation $w = \sum_{i=1}^n \lambda_i k(\cdot, x_i)$.

(5 Punkte)

**H 3.** (Programming task - regularization parameter and cross validation revisited)

See the accompanying notebook.

(5 Punkte)

**H 4.** (Programming task - bone mineral density estimation)

See the accompanying notebook.

(5 Punkte)