Theoretical Exercises

Statistical Learning, 2009

Niels Richard Hansen
April 20, 2009

The following exercises are going to play a central role in the course *Statistical learning*, block 4, 2009. The exercises are all of a nature that is somewhere in between an ordinary exercise and the theory that will be covered by the lectures. The exercises are divided between the participants so that each exercise is solved by a group of 2 students, who will give a subsequent oral presentation of the solution.

The oral presentation should take approximately 2×45 minutes, and the presenters should be careful in the presentation to give sufficient background so that the other participants can follow the presentation of the solution(s).

The solution and the presentation of a theoretical exercise are, like the compulsory assignments, evaluated as passed/not-passed.

1

Principal Components

Let Σ denote a $p \times p$ symmetric, positive semidefinite matrix (a covariance matrix). The *principal components* are defined to be an orthogonal basis of eigenvectors for the matrix. In matrix notation this can be written as follows:

$$\Sigma = V \Lambda V^T$$

where V is an orthogonal $p \times p$ matrix, $V^T V = I$, and $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_p)$ is a diagonal $p \times p$ matrix with the corresponding eigenvalues. The columns of V constitute the orthogonal basis of principal components. The eigenvalues are real and non-negative due to the fact that Σ is symmetric and positive semidefinite. The usual convention is to organize the eigenvalues in Λ in decreasing order. Note that if some of the eigenvalues are equal (for instance, if Σ does not have full rank and two or more of the eigenvalues are 0), then the orthogonal basis and thus the principal components are not unique.

If X denotes a p-dimensional random variable with covariance matrix Σ, the variance of a linear combination of the entries in X is

$$\mathbb{V}(\sum_{i=1}^{p} \beta_i X_i) = \mathbb{V}(\beta^T X) = \beta^T \Sigma \beta$$

where $\beta \in \mathbb{R}^p$ is any p-dimensional vector.

Let the columns of V (the orthogonal basis of $\mathbb{R}^p$ of principal components) be denoted $v_1, \ldots, v_p$. Let

$$H_q = \text{span}\{v_1, v_2, \ldots, v_q\}$$

denote the subspace of $\mathbb{R}^p$ spanned by the first (those with largest eigenvalues) q principal components for $q = 1, 2, \ldots, p$.

Question 1.1. *Show that*

$$\lambda_q = v_q^T \Sigma v_q = \max_{\substack{\beta \in H_{q-1}^{\perp} \\ \beta^T \beta = 1}} \beta^T \Sigma \beta. \tag{1.1}$$

Here $H_{q-1}^{\perp}$ denotes the orthogonal complement of H_{q-1} in $\mathbb{R}^p$ ($H_0^{\perp} = \mathbb{R}^p$).

This gives a sequence of linear combinations $v_1^T X, v_2^T X, \ldots, v_p^T X$.

Question 1.2. *Explain that the result above shows that $v_q^T X$ is the (unit norm) linear combination of the rows in X with the maximal variance subject to the constraint that $v_q^T X$ is uncorrelated with the variables $v_1^T X, \ldots, v_{q-1}^T X$.*

Consider next the situation that $\mathbf{X}$ is an $N \times p$ matrix of N repeated observations of X. The standard estimate of the expectation is

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

where x_i denotes that i'th row in $\mathbf{X}$. The standard estimate of the covariance matrix is

$$\hat{\Sigma} = \frac{1}{N-1} (\mathbf{X} - 1\hat{\mu}^T)^T (\mathbf{X} - 1\hat{\mu}^T)$$

with 1 denoting the column vector of ones. As it is common, to ease notation we will simply assume in the following that $\mathbf{X}$ has been centered so that the average of each column is 0 (formally, $\mathbf{X}$ has been replaced by $(\mathbf{X} - 1\hat{\mu}^T)$). Then the estimate of Σ is

$$\hat{\Sigma} = \frac{1}{N-1} \mathbf{X}^T \mathbf{X}.$$

The *estimated* principal components and corresponding eigenvalues are then the principal components and eigenvalues for the estimated covariance matrix $\hat{\Sigma}$.

Let in the following $K_q \subseteq \mathbb{R}^p$ denote a q-dimensional subspace of $\mathbb{R}^p$. The *rank-q-reconstruction error* of the data in $\mathbf{X}$ is defined as

$$\min_{K_q} \sum_{i=1}^{N} \min_{z \in K_q} ||x_i - z||^2.$$

The rank-q-reconstruction error is given in terms of a q-dimensional subspace that minimizes the total sums of squared distances from the observations to the subspace. The $z \in K_q$ that minimizes $||x_i - z||^2$ is the orthogonal projection onto K_q.

Question 1.3. *Fix K_q and let $w_1, \ldots, w_q$ be an orthonormal basis for K_q. Show that*

$$\sum_{i=1}^{N} \min_{z \in K_q} ||x_i - z||^2 = \sum_{i=1}^{N} ||x_i||^2 - (N-1) \sum_{j=1}^{q} w_j^T \hat{\Sigma} w_j.$$

Question 1.4. *Show that the rank-q-reconstruction error is obtained by taking $K_q = \hat{H}_q = \text{span}\{\hat{v}_1, \ldots, \hat{v}_q\}$ where $\hat{v}_1, \ldots, \hat{v}_p$ are the q first principal components for the estimated covariance matrix $\hat{\Sigma}$. Compute the reconstruction error in terms of the eigenvalues.*

This means that the best q-dimensional representation of N p-dimensional data vectors is given by projecting the observations onto the first q principal components of the estimated covariance matrix. From the principal components for the $\mathbf{X}$-matrix we get *principal components regression* where we regress an additional vector $\mathbf{y}$ of real observations on $\mathbf{X}[\hat{v}_1 \dots \hat{v}_q]$. This means that we take the q first principal components and form the corresponding q linear combinations of the columns in $\mathbf{X}$.

> **Question 1.5.** *Show that the resulting q columns are orthogonal vectors in $\mathbb{R}^N$. Explain how orthogonality implies that we can regress $\mathbf{y}$ by regression on each of the columns separately.*

Principal components regression reduce the dimensionality of the covariates (from p to q) by exploiting the internal distribution of the covariates. In particular, with $q = 1$ we get a one-dimensional projection of the covariates and a corresponding regression of $\mathbf{y}$.

For the final two questions we make the distributional assumption that conditionally on $\mathbf{X}$ the coordinates $y_1, \dots, y_N$ are independent all with variance σ^2.

> **Question 1.6.** *Let $\hat{\gamma}_\beta$ denote the estimated coefficient when we regress $\mathbf{y}$ on the column $\mathbf{X}\beta$ for some unit norm β. Show that*
>
> $$\min_{\beta : ||\beta|| = 1} \mathbb{V}(\hat{\gamma}_\beta) = \mathbb{V}(\hat{\gamma}_{\hat{v}_1}).$$

Thus the first principal component provides the one-dimensional projection of the covariates where the regression coefficient has the least variance. We could instead ask for the one-dimensional projection that has the largest correlation with $\mathbf{y}$.

> **Question 1.7.** *Show that the empirical correlation,*
>
> $$\hat{Corr}(Y, X\beta) = \frac{y^T \mathbf{X}\beta}{\sqrt{\mathbf{y}^T \mathbf{y} \beta^T \mathbf{X}^T \mathbf{X}\beta}},$$
>
> *is maximized by taking $\beta = \hat{\beta}^{ls}$ – the ordinary least squares regression estimate.*

2

Ridge Regression

If $\mathbf{X}$ denotes an $N \times p$ matrix of N p-dimensional covariates and $\mathbf{y}$ denotes an N-dimensional vector of observations we consider the penalized residual sum of squares

$$\mathrm{RSS}_\lambda(\beta) = (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta) + \lambda\beta^T\beta$$

for a $\lambda > 0$. The *ridge regression* estimate of β is defined as the β that minimizes this penalized residual sum of squares. For $\lambda > 0$ there is always a unique solution.

As is often the case we will assume that the matrix of covariates as well as the observation vector have been centered, that is, the average of the columns as well as of $\mathbf{y}$ equal 0.

Question 2.1. *Solve Exercise 3.5 to show that the assumption above can be made.*

The ordinary least squares estimate is obtained by minimizing $\mathrm{RSS}_0(\beta)$, and the solution is only unique if $\mathbf{X}$ has rank p (note, since the columns have been centered, this requires at least $p + 1$ rows). Any least squares solution β^{ls} – unique or not - fulfills

$$\mathbf{X}^T\mathbf{X}\beta^{\mathrm{ls}} = \mathbf{X}^T y$$

and we introduce

$$t = \min_{\beta : \mathbf{X}^T\mathbf{X}\beta = \mathbf{X}^T y} \beta^T\beta.$$

Question 2.2. *With $\mathbf{X} = UDV^T$ the singular value decomposition of $\mathbf{X}$ show that if $\hat{\beta}(\lambda)$ is the minimizer of $RSS_\lambda(\beta)$ then*

$$\hat{\beta}(\lambda)^T\hat{\beta}(\lambda) = \sum_{i=1}^{p} \frac{d_i^2}{(d_i^2 + \lambda)^2} y^T u_i u_i^T y$$

where d_i, $i = 1, \ldots, p$ are the singular values and u_i, $i = 1, \ldots, p$ are the columns of U in the singular value decomposition.

Question 2.3. *Use the result above to show that* $\hat{\beta}(\lambda)^T\hat{\beta}(\lambda) < t$ *for* $\lambda > 0$ *and that the function*

$$\lambda \mapsto s(\lambda) := \hat{\beta}(\lambda)^T\hat{\beta}(\lambda)$$

is a continuous, strictly decreasing function whose limit for $\lambda \to \infty$ *equals 0.*

Thus s maps the interval $(0, \infty)$ in a one-to-one manner onto the interval $(0, t)$.

Question 2.4. *Show that the minimizer of* $RSS_\lambda(\beta)$ *is also a minimizer of*

$$(\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta)$$

subject to the constraint

$$\beta^T\beta \leq s(\lambda).$$

Question 2.5. *Conversely, show that a minimizer of*

$$(\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta)$$

subject to the constraint

$$\beta^T\beta \leq s(\lambda)$$

is also a minimizer of $RSS_\lambda(\beta)$*. Argue that the constraint minimization problem above yields the ordinary least squares estimate whenever* $s \geq t$*.*

The constraint minimization formulation provides the interpretation that the solution is simply a least squares estimate but on a restricted parameter space – restricted by the norm constraint $\beta^T\beta \leq s$ on the norm of the parameter vector β. Thus the penalization can essentially be regarded as a model restriction. Note, however, the subtle, data dependent relation between λ and s that makes the transformation between the two optimization problems data dependent. Thus, using the penalization formulation for a fixed λ the nature of the model restriction imposed by λ is *data dependent*. In practice, λ may even be determined from the data to optimize empirically the tradeoff between bias and variance for the ridge regression estimator.

We know that the predicted values without penalization are given as

$$\hat{y} = \mathbf{X}\hat{\beta}(0) = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}.$$

Question 2.6. *Show that for the projection* $P = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ *onto the column space of* $\mathbf{X}$ *we have* $\mathrm{trace}(P) = p$ *and* $P^2 = P$*.*

We also know that the predicted values with penalization are given as

$$\hat{y} = \mathbf{X}\hat{\beta}(0) = \mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda I)^{-1}\mathbf{X}^T\mathbf{y}.$$

Question 2.7. *Show that for the so-called smoother* $S_\lambda = \mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda I)^{-1}\mathbf{X}^T$ *we have*

$$\mathrm{trace}(S_\lambda) = \sum_{i=1}^{p} \frac{d_i^2}{d_i^2 + \lambda} < p.$$

and $S_\lambda^2 \prec S_\lambda$*.*

3

Finite Dimensional Reproducing Kernel Hilbert Spaces

Let $\Omega \subseteq \mathbb{R}^p$ and let $\varphi_i : \Omega \to \mathbb{R}$ be given functions for $i \in I$. Here I is a finite index set.

Define a *kernel* $K : \Omega \times \Omega \to \mathbb{R}$ in terms of the φ_i functions for $i \in I$ by

$$K(x, y) = \sum_{i \in I} \varphi_i(x)\varphi_i(y).$$

Define the *space of functions* $\mathcal{H}_K$ as the functions $f : \Omega \to \mathbb{R}$ where

$$f(x) = \sum_{i \in I} \beta_i \varphi_i(x)$$

for a sequence of real coefficients $\beta_i \in \mathbb{R}$. This space of functions can be seen as a finite dimensional vector space spanned by the functions φ_i for $i \in I$. To avoid redundancy in the representation of $f \in \mathcal{H}_K$ we will always assume that the φ_i functions are linearly independent.

On $\mathcal{H}_K$ we define *an inner product*

$$\langle f, g \rangle = \sum_{i \in I} \beta_i \delta_i$$

where $f = \sum_{i \in I} \beta_i \varphi_i$ and $g = \sum_{i \in I} \delta_i \varphi_i$.

Denote by $K(\cdot, y)$ and $K(x, \cdot)$ the functions

$$x \mapsto K(x, y)$$

for fixed $y \in \Omega$ and

$$y \mapsto K(x, y)$$

for fixed $x \in \Omega$, respectively.

Question 3.1. *Show that for any $y \in \Omega$ we have $K(\cdot, y) \in \mathcal{H}_K$ and that for any $f \in \mathcal{H}_K$ we have*

$$\langle K(\cdot, y), f \rangle = f(y).$$

Then show that for any $x \in \Omega$ we also have $K(x, \cdot) \in \mathcal{H}_K$ and that

$$\langle K(\cdot, y), K(x, \cdot) \rangle = K(x, y).$$

This last property is known as *the reproducing property* of the space of functions $\mathcal{H}_K$ and the kernel K.

The inner product space $\mathcal{H}_K$ is an example of a finite dimensional Hilbert space. The norm on this space is defined in terms of the inner product as

$$||f||_K^2 = \langle f, f \rangle = \sum_{i \in I} \beta_i^2$$

when $f = \sum_{i \in I} \beta_i \varphi_i$. Such a Hilbert space is known as a *reproducing kernel Hilbert space* due to the fact that the inner product and the kernel "play together" to give the reproducing property as shown above.

The kernel plays a central role if we want to estimate the coefficients in an expansion of $f \in \mathcal{H}_K$ from (noisy) data. To estimate the parameters we will minimize the empirical loss for a loss function $L : \mathbb{R} \times \Omega \to [0, \infty)$ but with a ridge regression type of penalty term added.

If we observe $(y_1, x_1), \ldots, (y_N, x_N)$ with $x_1, \ldots, x_N \in \mathbb{R}^p$ and $y_1, \ldots, y_N \in \mathbb{R}$ we will minimize

$$\sum_{i=1}^N L(y_i, f(x_i)) + \lambda ||f||_K^2$$

over $f \in \mathcal{H}_K$ and for $\lambda \geq 0$ a given constant. Equivalently, we can minimize

$$\sum_{i=1}^N L(y_i, \sum_{j \in I} \beta_j \varphi_j(x_i)) + \lambda \sum_{j \in I} \beta_j^2$$

over all sequences $(\beta_j)_{j \in I}$ of real value coefficients. We see the resemblance to ridge regression in the penalty term that also in this case involves the sum of the squared coefficients.

If $x_1, \ldots, x_N \in \mathbb{R}^p$ we define the matrix

$$\mathbf{K} = \{K(x_i, x_j)\}_{i,j=1,\ldots,N}.$$

We assume that the x_i's are all different.

Question 3.2. *Show that $\mathbf{K}$ is a symmetric, positive semidefinite matrix. When is it positive definite?*

Question 3.3. *Let* $f = \sum_{i=1}^N \alpha_i K(\cdot, x_i)$. *Show that* $f \in \mathcal{H}_K$ *and that*

$$||f||_K^2 = \alpha^T \mathbf{K} \alpha.$$

Then show that if $\rho \in \mathcal{H}_K$ *is a function orthogonal to the functions* $K(\cdot, x_i)$ *for* $i = 1, \ldots, N$, *that is,*

$$\langle \rho, K(\cdot, x_i) \rangle = \rho(x_i) = 0$$

for $i = 1, \ldots, N$, *then*

$$\sum_{i=1}^N L(y_i, f(x_i) + \rho(x_i)) + \lambda ||f + \rho||_K^2 \geq \sum_{i=1}^N L(y_i, f(x_i)) + \lambda ||f||_K^2$$

with equality if and only if $\rho = 0$ *(*ρ *is constantly equal to 0).*

Question 3.4. *Show that the minimizer of*

$$\sum_{i=1}^N L(y_i, f(x_i)) + \lambda ||f||_K^2$$

is of the form

$$f(x) = \sum_{i=1}^N \alpha_i K(x, x_i)$$

for some $\alpha_i \in \mathbb{R}$, $i = 1, \ldots, N$, *and show that it can be found by minimizing*

$$L(\mathbf{y}, \mathbf{K}\alpha) + \lambda \alpha^T \mathbf{K} \alpha \tag{3.1}$$

where we use vector notation and let $L(\mathbf{y}, \mathbf{z}) = \sum_{i=1}^N L(y_i, z_i)$ *for vectors* $\mathbf{y} = (y_i)_{i=1,\ldots,N}$ *and* $\mathbf{z} = (z_i)_{i=1,\ldots,N}$.

Question 3.5. *If* L *is the squared error loss function,* $L(y, z) = (y - z)^2$, *show that the minimizer of (3.1) is unique and given as*

$$\hat{\alpha} = (\mathbf{K} + \lambda I)^{-1} \mathbf{y}.$$

Question 3.6. *Show that with* $\Omega = \mathbb{R}^p$,

$$K(x, y) = (1 + \sum_{i=1}^p x_i y_i)^d$$

is a kernel that is given as a sum of $\binom{p+d}{d}$ *polynomials (the* φ_i *functions).*

The major point in reproducing kernel Hilbert spaces is that the number of basis functions, as in the polynomial example above, may we very large – much larger than N, the number of observations. If the kernel, on the other hand, is easy to evaluate we can typically solve the minimization problem over $\mathcal{H}_K$ anyway using the kernel formulation in (3.1), and the solution is expressed entirely in terms of kernel evaluations.

A useful, but mathematically more advanced generalization involves *infinite* dimensional reproducing kernel Hilbert spaces. Starting with a kernel $K : \Omega \times \Omega \to \mathbb{R}$ with the property that the matrix $\mathbf{K}$, as defined above, for *any* dataset $x_1, \ldots, x_N \in \mathbb{R}^p$

and any $N \geq 1$, is symmetric and positive definite the *Moore-Aronszajn* theorem provides a completely abstract construction of a corresponding reproducing kernel Hilbert space. The corresponding Hilbert space may be finite dimensional, as the polynomial example above, or infinite dimensional. The existence of an expansion of the kernel as an (infinite) sum of φ_i-functions is, however, a more delicate matter. Under certain regularity conditions the various versions of *Mercer's theorem* provide explicit as opposed to abstract Hilbert spaces in combination with such an expansion. Note that the literature can be confusing on this point where Mercer's theorem is often dragged into the discussion when it is not needed and references to Mercer's theorem obscure rather than clarify the picture. We need the reproducing kernel Hilbert space and the kernel, but the existence of an expansion in terms of φ_i-functions is actually not needed in general. We used it here in the finite dimensional case solely to be able to construct the Hilbert spaces explicitly.

There is one catch, which seems most easily understood in terms of finite or infinite expansions in terms of φ_i-functions. Nice kernels that are easy to evaluate define implicitly the φ_i-functions and in particularly the scale of these functions. If we want to scale these basis functions in a different way, e.g. to penalize certain properties more or less, we would typically loose the efficient evaluation of the kernels. Thus the price we pay for the easy solution of a minimization problem is that we can not control the scales of the basis functions and we just have to live with those imposed by the kernel.

4

Penalized Logistic Regression

We have a dataset $(y_1, x_1), \ldots, (y_N, x_N)$ with the observations $y_i \in \{0, 1\}$ and the covariates $x_i \in \mathbb{R}$. We assume that the observations are realizations of iid variables having the same distribution as (Y, X), where the conditional probability, $p(x)$, of $Y = 1$ given $X = x$ is given as

$$\text{logit}(p(x)) = f(x),$$

or alternatively

$$p(x) = P(Y = 1|X = x) = \frac{\exp(f(x))}{1 + \exp(f(x))}.$$

Question 4.1. *Show that the minus-log-likelihood function, as a function of the "parameter" f is given as*

$$l(f) = \log(1 + \exp(f(x_i))) - \sum_{i=1}^{N} y_i f(x_i),$$

where $f : \mathbb{R} \to \mathbb{R}$ is some function (the conditional log-odds).

Assume then that f is given by a finite basis expansion

$$f_\beta(x) = \sum_{j=1}^{p} \beta_j \varphi_j(x)$$

with the φ_j-functions known and fixed and $\beta \in \mathbb{R}^p$. Write $p_\beta(x) = \frac{\exp(f_\beta(x))}{1+\exp(f_\beta(x))}$, let $\mathbf{p}_\beta(x)$ denote the N-dimensional vector of $p_\beta(x_i)$, and introduce the β-dependent *weights*

$$w_i(\beta) = p_\beta(x_i)(1 - p_\beta(x_i)).$$

Let $\mathbf{X}$ denote the $N \times p$ matrix with the (i, j)'th entry being $\varphi_j(x_i)$.

Question 4.2. *Show that the first derivative of l as a function of β is*

$$D_\beta(l)(\beta) = (\mathbf{p}_\beta(x) - \mathbf{y})^T \mathbf{X}$$

and that the second derivative is

$$D^2_\beta l(\beta) = \mathbf{X}^T \mathbf{W}(\beta) \mathbf{X}.$$

where $\mathbf{W}(\beta) = diag(w_1(\beta), \ldots, w_N(\beta))$ is the $N \times N$ diagonal matrix with the weights $w_i(\beta)$ in the diagonal.

Question 4.3. *Show that a single Newton-Raphson step*[a] *can be written as*

$$\beta^1 = (\mathbf{X}^T \mathbf{W}(\beta^0) \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}(\beta^0) z(\beta^0)$$

where

$$z(\beta^0) = \mathbf{X}\beta^0 + \mathbf{W}(\beta^0)^{-1}(\mathbf{y} - \mathbf{p}_{\beta^0}(x)).$$

[a] The solution to $D_\beta(l)(\beta^0) + (\beta - \beta^0)^T D^2_\beta l(\beta^0) = 0$

The entries in the vector $z(\beta^0)$ are sometimes called the *adjusted responses*. One can view the Newton-Raphson algorithm as a two-step procedure, where we first compute the adjusted responses based on the given parameter vector β^0, and then solve the following weighted least squares problem; minimize

$$(z(\beta^0) - \mathbf{X}\beta)^T \mathbf{W}(\beta^0)(z(\beta^0) - \mathbf{X}\beta),$$

where the weight matrix $\mathbf{W}(\beta^0)$ also depends upon β^0.

Consider instead the *penalized minus-log-likelihood* as a function of β;

$$l^1(\beta; \lambda) = l(\beta) + \frac{\lambda}{2} \beta^T \mathbf{\Omega} \beta$$

for $\lambda > 0$ and $\mathbf{\Omega}$ a fixed $p \times p$ matrix.

Question 4.4. *Show that a step in the Newton-Raphson algorithm for maximizing $l^1(\beta; \lambda)$ for fixed λ can be computed in the same way as above except from the fact that $\mathbf{X}^T \mathbf{W}(\beta^0) \mathbf{X}$ is replaced by $\mathbf{X}^T \mathbf{W}(\beta^0) \mathbf{X} + \lambda \mathbf{\Omega}$. That is, the adjusted response, $z(\beta^0)$, is computed the same way, but then*

$$\beta^1 = (\mathbf{X}^T \mathbf{W}(\beta^0) \mathbf{X} + \lambda \mathbf{\Omega})^{-1} \mathbf{X}^T \mathbf{W}(\beta^0) z(\beta^0).$$

We return to the general setup without a priori restrictions on f. We assume that the x_i's are all different and are all in the interval $[a, b]$. We consider the *penalized minus-log-likelihood*

$$l^2(f; \lambda) = l(f) + \frac{\lambda}{2} \int_a^b (f''(x))^2 \mathrm{d}x$$

over the set of twice differentiable functions f.

Question 4.5. *Show that, for fixed λ, the minimum of $l^2(f;\lambda)$ over the the set of twice differentiable functions on $[a,b]$ is attained for a function*

$$f(x) = \sum_{j=1}^{N} \beta_j \varphi_j(x)$$

where φ_j for $j = 1, \ldots, N$ constitute a basis for the set of natural cubic splines with knots in the x_i's. Consult Exercise 5.7 for the similar problem but with quadratic loss.

Question 4.6. *Show that the resulting (finite-dimensional) optimization problem consists of maximizing $l^1(\beta;\lambda)$ with $\mathbf{\Omega}$ the $N \times N$ matrix with entries*

$$\Omega_{ij} = \int_a^b \varphi_i''(x)\varphi_j''(x)\mathrm{d}x.$$

Argue that a specific choice of a and b does not matter as long as the x_i's are contained in the interval $[a,b]$ and that one can take a and b to be the minimum and maximum of the x_i's, respectively.

5

Support Vector Machines

In classification problems with two classes it is common to label the classes as ± 1. Thus we consider an observation $y \in \{-1, 1\}$ and a covariate vector $x \in \mathbb{R}^p$. If

$$f : \mathbb{R}^p \to \mathbb{R}$$

is any function we can construct a classification procedure by predicting the value of y for given x as $\text{sign}(f(x))$. If $p(x)$ denotes the conditional probability of $y = 1$ given x the Bayes classifier is given in this way in terms of the logit transform of $p(x)$, that is

$$f(x) = \text{logit}(p(x)) = \log \frac{p(x)}{1 - p(x)}.$$

An arbitrary function produces a correct prediction if $yf(x) > 0$ and a wrong prediction if $yf(x) < 0$. Using a zero-one loss function for estimation of f generally leads to none-uniqueness and a quite awful optimization problem. Other loss functions used in the literature are the squared error loss

$$L(y, f(x)) = (y - f(x))^2 = (1 - yf(x))^2,$$

and the binomial minus-log-likelihood (where we thus regard f to be the logit transform of $p(x)$)

$$L(y, f(x)) = \log(1 + \exp(-yf(x))).$$

A third sensible loss function is defined by

$$L(y, f(x)) = (1 - yf(x))_+$$

where the subscript "+" indicates positive part, that is, $z_+ = \max\{z, 0\}$.

Assume that $h : \mathbb{R}^p \to \mathbb{R}^M$ is a fixed function, and let

$$f(x) = \beta^T h(x) + \beta_0$$

for $\beta \in \mathbb{R}^M$ and $\beta_0 \in \mathbb{R}$. Let $(y_1, x_1), \ldots, (y_N, x_N)$ be a dataset and consider the following penalized loss function

$$\sum_{i=1}^{N} (1 - y_i f(x_i))_+ + \lambda \beta^T \beta \tag{5.1}$$

with $\lambda > 0$. The penalty function is of the same type as in ridge regression, and we seek to minimize (5.1). If we changed the loss function to the squared error loss, the minimization would indeed be a ridge regression problem with its explicit solution.

Question 5.1. *Show that the function given by (5.1) is strictly convex and has a unique minimum.*

Question 5.2. *Show that minimization of (5.1) is equivalent to minimizing*

$$\beta^T \beta + \frac{1}{\lambda} \sum_{i=1}^{N} \xi_i$$

subject to the constraints

$$\xi_i \geq 0, \quad y_i(\beta^T h(x_i) + \beta_0) \geq 1 - \xi_i$$

for $i = 1, \ldots, N$.

The latter minimization problem is a quadratic optimization problem with linear constraints – something that can be solved numerically.

If $h(x) = (\varphi_1(x), \varphi_2(x), \ldots, \varphi_M(x))$ we may observe that

$$f(x) = \beta_0 + \sum_{i=1}^{M} \beta_i \varphi_i(x)$$

so we are working with a setup (almost) like in the exercise on reproducing kernel Hilbert spaces, where the function we estimate is given as a linear combination of basis functions. The corresponding kernel reads

$$K(x, y) = \sum_{i=1}^{M} \varphi_i(x) \varphi_i(y) = h(x)^T h(y),$$

and we let

$$\mathbf{K} = \{K(x_i, x_j)\}_{i,j=1,\ldots,N} = \{h(x_i)^T h(x_j)\}_{i,j=1,\ldots,N}.$$

Question 5.3. *Show, using the results on reproducing kernel Hilbert spaces, that the minimizer of (5.1) can be given as*

$$f(x) = \beta_0 + \sum_{i=1}^{N} \alpha_i K(x, x_i)$$

for $\beta_0, \alpha_i \in \mathbb{R}$, $i = 1, \ldots, N$, *and show that minimization of (5.1) is equivalent to minimization of*

$$\sum_{i=1}^{N} (1 - y_i(\beta_0 + [\mathbf{K}\alpha]_i))_+ + \lambda \alpha^T \mathbf{K} \alpha. \tag{5.2}$$

Question 5.4. *Show that minimization of (5.2) is equivalent to minimizing*

$$\alpha^T \mathbf{K}\alpha + \frac{1}{\lambda}\sum_{i=1}^{N} \xi_i$$

subject to the constraints

$$\xi_i \geq 0, \quad y_i(\beta_0 + [\mathbf{K}\alpha]_i) \geq 1 - \xi_i$$

for $i = 1, \ldots, N$.

The last result is an example of the so-called *kernel property*, that a potentially high-dimensional (in general, even infinite dimensional) optimization problem can be reduced to a manageable optimization problem with the number of parameters being of the size of the dataset. The crux of the matter is the ability to efficiently evaluate the kernel, thus it is the implicit structure of the basis expansion of nice kernels that determines how the penalization affects the resulting estimate.

In the following 3 questions you are going to derive a result which can be obtained as a consequence of general results from the theory of constraint optimization. However, you can derive these results here from first principle without the knowledge of anything but classical calculus.

Question 5.5. *Show that if* (α_0, ξ) *is a solution to the minimization problem then either* $\xi_i = 0$ *or* $\xi_i = 1 - y_i(\beta_0 + [\mathbf{K}\alpha_0]_i)$.

In the following we let (α_0, ξ) denote a solution. Due to the result above the following three sets provide a partition of the indices $\{1, \ldots, N\}$.

$$\begin{aligned}
\mathcal{A} &= \{i \mid \xi_i = 1 - y_i(\beta_0 + [\mathbf{K}\alpha_0]_i) > 0\} \\
\mathcal{B} &= \{i \mid \xi_i = 0, 1 - y_i(\beta_0 + [\mathbf{K}\alpha_0]_i) > 0\} \\
\mathcal{C} &= \{i \mid \xi_i = 1 - y_i(\beta_0 + [\mathbf{K}\alpha_0]_i) = 0\}
\end{aligned}$$

Define also

$$\begin{aligned}
C &= \{\alpha \mid 1 - y_i(\beta_0 + [\mathbf{K}\alpha]_i) = 0,\ i \in \mathcal{C}\} \\
O &= \{\alpha \mid 1 - y_i(\beta_0 + [\mathbf{K}\alpha]_i) > 0,\ i \in \mathcal{A} \cup \mathcal{B}\}.
\end{aligned}$$

Finally we define the vector $\tilde{\mathbf{y}} \in \mathbb{R}^N$ by

$$\tilde{y}_i = \begin{cases} y_i & \text{if} \quad i \in \mathcal{A} \\ 0 & \text{if} \quad i \notin \mathcal{A} \end{cases}$$

Question 5.6. *Argue that* O *is open. Define*

$$\mathcal{L}(\alpha) = \alpha^T \mathbf{K}\alpha - \frac{1}{\lambda}\tilde{\mathbf{y}}^T \mathbf{K}\alpha - \frac{1}{\lambda}\sum_{i \in \mathcal{A}} 1 - \beta_0 y_i.$$

Show that $\mathcal{L}(\alpha_0) \leq \mathcal{L}(\alpha)$ *for* $\alpha \in C \cap O$.

Define $\Lambda = \{\lambda \in \mathbb{R}^N \mid \lambda_i = 0,\ i \notin \mathcal{C}\}$ and $V = \{\mathbf{K}\lambda \mid \lambda \in \Lambda\}$.

> **Question 5.7.** *Show that if $\rho \in V^\perp$ then $\alpha_0 + \varepsilon\rho \in O \cap C$ for some $\varepsilon > 0$. Use this to show that due to fact that α_0 is a minimizer in $C \cap O$ we have $\nabla\mathcal{L}(\alpha_0) \in V$ and conclude that*
>
> $$\alpha_0 = \frac{2}{\lambda}\tilde{\mathbf{y}}^T + \lambda$$
>
> *for some $\lambda \in \Lambda$. Conclude in particular that $\alpha_{0,i} = 0$ for $i \in \mathcal{B}$.*

We call the points x_i with $\alpha_{0,i} > 0$ the *support vectors*. We can see from the expansion of f in terms of the kernel that in reality we only need to expand f in terms of the kernel evaluated in the support vectors. The result above shows that only the so-called *active constraints* for the solution contribute with a support vector. A sparse solution (many zeros in the α vector) is desirable as this will result in faster evaluations of the estimated f when we want to apply f for predictions.

6

Linear Smoothers and Cross-Validation

If X is p-dimensional, real random variable, Y a one dimensional, real random variable and $f : \mathbb{R}^p \to \mathbb{R}$ is a predictor of Y given X, a quantity of interest is the expected prediction error

$$\mathrm{EPE}(f) = \mathbb{E}(L(Y, f(X)))$$

for $L : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ the loss function. If $f = \hat{f}$ is estimated based on the dataset $(x_1, y_1), \ldots, (x_N, y_N)$ then

$$\mathrm{EPE}(\hat{f}) = \mathbb{E}(L(Y, f(X))|X_1 = x_1, \ldots, X_n = x_n, Y_1 = y_1, \ldots, Y_n = y_1)$$

is the expected prediction error for the estimated predictor conditionally on the data. The *training error*

$$\mathrm{err}(y_1, \ldots, y_n, x_1, \ldots, x_n) = \frac{1}{N} \sum_{i=1}^{N} L(y_i, \hat{f}(x_i)),$$

which is the average loss over the same dataset that was used for estimating $\hat{f}$ will generally underestimate the expected prediction error for $\hat{f}$. A serious problem is that the test error will typically be monotonely decreasing as a function of model complexity and is therefore not of much use in model selection. We assume here that the pairs (x_i, y_i) are realization of N iid random variables.

A partial solution to the problem with the test error is to introduce the *in-sample error*

$$\mathrm{Err}_{\mathrm{in}} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}(L(Y_i^{\mathrm{New}}, \hat{f}(x_i))|X_1 = x_1, \ldots, X_n = x_n),$$

which is the average expected loss over the x_i's in the dataset. Note, the expectation is over the Y_i's and the new, independent variable Y_i^{New}, whose distribution is *the conditional distribution of* Y_i *given* $X_i = x_i$. The *optimism* in the training error is defined as

$$\text{op} = \text{Err}_{\text{in}} - \mathbb{E}\left(\text{err}(Y_1, \ldots, Y_n, x_1, \ldots, x_n)\right).$$

Question 6.1. *Show that with* $L(y, z) = (y - z)^2$ *the squared error loss,*

$$op = \frac{2}{N} \sum_{i=1}^{N} Cov(\hat{f}(x_i), Y_i)$$

where $\hat{f}(x_i)$ *is a random variable as a function of* $Y_1, \ldots, Y_N$.

A *linear smoother* is an estimating procedure where the N-dimensional vector $\hat{\mathbf{f}}$ of fitted values $\hat{f}(x_i)$ for $i = 1, \ldots, N$ is given by

$$\hat{\mathbf{f}} = \mathbf{S}\mathbf{y}$$

for a matrix $\mathbf{S}$ not depending upon the y_i observations.

Question 6.2. *Show that if the conditional variance*

$$\sigma^2 = \mathbf{V}(Y|X = x)$$

does not depend[a] *upon* x *then for a linear smoother it holds that*

$$\sum_{i=1}^{N} Cov(\hat{f}(x_i), Y_i) = trace(\mathbf{S})\sigma^2.$$

[a] As if $Y = f(X) + \varepsilon$ with ε a random variable independent of X.

Question 6.3. *Show that if* $\hat{\sigma}^2$ *is an unbiased estimator of* σ^2 *then*

$$\hat{Err}_{in} = err + \frac{2}{N} trace(\mathbf{S})\hat{\sigma}^2$$

is an unbiased estimator of the in-sample error Err_{in} *when using a linear smoother given by the smoother matrix* $\mathbf{S}$.

The estimate above offers a quantitative compromise – a small training error is typically obtained by a smoother where trace($\mathbf{S}$) is large.

An alternative to the in-sample error is to introduce the *test* or *generalization error*

$$\text{Err} = \mathbb{E}(L(Y, \hat{f}(X)).$$

The difference from the definition of the expected prediction error above is that when defining the generalization error we take expectation over (X, Y) *as well* as the variables $(X_1, Y_1), \ldots, (X_N, Y_N)$ that enter in the computation of $\hat{f}$. The typical estimate of the generalization error is constructed via *cross-validation*. The general cross-validation procedure consists of splitting the dataset into groups (K groups in K-fold cross-validation) and then estimate f on the basis of $K - 1$ groups and

estimate the prediction error on the basis of the remaining group. This process is repeated – leaving out all groups for the estimation one at a time. The special case is *leave-one-out* cross-validation or N-fold cross-validation. With $\hat{f}^{-i}$ denoting the estimated predictor based on all observations except the i'th (x_i, y_i).

The leave-one-out CV-estimate of Err is defined as

$$\hat{\text{Err}} = \frac{1}{N}\sum_{i=1}^{N} L(y_i, \hat{f}^{-i}(x_i)).$$

We consider the linear smoother given by the smoother matrix $\mathbf{S}$. To come up with convenient formulas for the leave-one-out CV-estimate we need the smoother for the whole dataset to be related to the smoother for a dataset where we remove one data point.

Question 6.4. *Show that for a linear smoother with*

$$\hat{f}^{-i}(x_i) = \sum_{j=1, j\neq i}^{N} \frac{S_{ij}}{1 - S_{ii}} y_j \tag{6.1}$$

it holds that

$$y_i - \hat{f}^{-i}(x_i) = \frac{y_i - \hat{f}(x_i)}{1 - S_{ii}}$$

Question 6.5. *Explain why the result above can be used to compute the estimate $\hat{Err}$ efficiently if we use the squared error loss – if we can compute S_{ii} efficiently.*
Explain that (6.1) can be understood as follows: If we estimate $\hat{f}^{-i}$ based on the dataset excluding (y_i, x_i) and subsequently include the data point $(\hat{f}^{-i}(x_i), x_i)$ – using the predicted value of y_i, $\hat{f}^{-i}(x_i)$, instead of y_i – then the prediction of y_i based on this enlarged dataset is still $\hat{f}^{-i}(x_i)$.

Question 6.6. *Show that ordinary least squares regression, Ridge regression, and cubic spline fits are all linear smoothers that fulfill (6.1).*

For some smoothers it is computationally easier to compute trace($\mathbf{S}$) than to compute the diagonal elements themselves. Approximating all diagonal elements by the constant trace($\mathbf{S}$)/N (pretending that all diagonal elements are equal) leads to the so-called *generalized*[1] *cross-validation* estimate

$$\text{GCV} = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{y_i - \hat{f}(x_i)}{1 - \text{trace}(\mathbf{S})/N}\right]^2.$$

Question 6.7. *Show by a second order Taylor expansion of $(1-x)^{-2}$ that if trace($\mathbf{S}$) $\ll N$ then*

$$\left[\frac{y_i - \hat{f}(x_i)}{1 - \mathit{trace}(\mathbf{S})/N}\right]^2 \simeq (y_i - \hat{f}(x_i))^2(1 + \frac{2}{N}\mathit{trace}(\mathbf{S})).$$

Explain how this gives a relation between GCV and the estimate $\hat{Err}_{in}$ of the in-sample error.

[1] It seems that "approximate" would have been are better choice than "generalized".