BISS: Regularization Methods for High Dimensional Learning

Due: 30/6/2012 or 31/12/2012.

Note: there are 8 problems and a total of 46 points.

Problem 1 [**Points: 4**] One common preprocessing in machine learning is to center the data. In this problem we will see how this can be related to working with an (unpenalized) off-set term in the solution. Consider the usual Tikhonov regularization with a linear kernel, but assume that there is an unpenalized offset term b,

$$\min_{w\in\mathbb{R}^d,\, b\in\mathbb{R}} \left\{ \frac{1}{n}\sum_{i=1}^{n} \big(\langle w, x_i\rangle + b - y_i\big)^2 + \lambda\|w\|^2 \right\}$$

and let (w^*, b^*) be the solution of the above problem.

For $i = 1, \ldots, n$, denote by $x_i^c = x_i - \bar{x}$, $y_i^c = y_i - \bar{y}$ the centered data, where $\bar{y}, \bar{x}$ are the output and input means respectively. Show that w^* also solves

$$\min_{w\in\mathbb{R}^d} \left\{ \frac{1}{n}\sum_{i=1}^{n} \big(\langle w, x_i^c\rangle - y_i^c\big)^2 + \lambda\|w\|^2 \right\}. \tag{1}$$

and determine b^*.

Problem 2 In this problem we will utilize the concept of feature map.

(a) [**Points: 2**] The distance between two elements $\Phi(x), \Phi(s)$ of a feature space induced by some kernel K can be seen as a new distance $d(x, x')$ in the input space. Show that such a distance can always be calculated without knowing the explicit form of the feature map itself.

(b) [**Points: 4**] You are given a dataset of x, y pairs $\{(x_i, y_i)\}_{i=1}^N$, with $x_i \in X$ and $y_i \in \{-1, 1\}$. Assume that n_+, n_- of the x_i have label $+1, -1$, respectively (so $n_+ + n_- = N$), and let's also assume that we are given a kernel K and an associated feature map $\Phi : X \to \mathcal{F}$ satisfying

$$K(x, x') = \langle \Phi(x), \Phi(x')\rangle_{\mathcal{F}}.$$

Derive a classification rule, involving only kernel products (and the sign function), that assigns to a new test point the label of the class whose mean is closest *in the feature space.*

Problem 3 In (binary) classification problems one aims at finding a classification rule (also called the "decision rule") which is a binary valued function on the input space $c : X \to$

$\{1, -1\}$. The quality of a classification rule can be naturally measured by means of the so called misclassification error defined by

$$R(c) = \mathbb{P}\{c(x) \neq y\}.$$

If we introduce the misclassification loss $V(c(x), y) = \theta(-yc(x))$, where $\theta(s) = 1$ if $s > 0$ and $\theta(s) = 0$ otherwise, the misclassification error can be rewritten as

$$R(c) = \int_{X \times Y} \theta(-yc(x))p(x)p(y|x)dxdy.$$

Direct minimization of the misclassification error is not computationally feasible mostly because the misclassification loss is not convex. In practice, one usually looks for real valued (rather than binary valued) functions $f : X \to \mathbf{R}$ and replaces $\theta(-yc(x))$ with some convex loss $V(-yf(x))$. A classification rule is then obtained by taking the sign, that is $c(x) = \text{sign}(f(x))$. Commonly chosen loss functions are the hinge loss and square loss (see class). Note that in this case the error is measured by the expected error

$$I[f] = \int_{X \times Y} V(-yf(x))p(x)p(y|x)dxdy.$$

However, there is still the problem of relating the convex approximation to the original classification problem.

With the above discussion in mind, answer the following questions:

(a) **[Points: 2]** The minimizer of $R(c)$ over all possible decision rules is the so called Bayes decision rule $b : X \to \{1, -1\}$. Find the explicit form of b.
(hint: Consider $R(c|x) = \int_Y \theta(-yc(x))p(y|x)dy$ and calculate $b(x)$ point-wise.)

(b) **[Points: 2]** Calculate the explicit form of the minimizer of $I[f]$ if V is the hinge loss. (hint: use a trick similar to the previous hint) How is it related to Bayes decision rule?

(c) **[Points: 2]** Check that the square loss can be written as $V(-yf(x))$. Calculate the explicit form of the minimizer of $I[f]$ if V is the square loss. How is it related to Bayes decision rule?

Problem 4 Consider a bounded loss function $V : \mathbb{R} \times \mathbb{R} \to (0, M]$ and a hypothesis space comprised of N distinct functions, $\mathcal{H} = \{f_1, \ldots, f_N\}$.

(a) **[Points: 2]** Prove that for all $\epsilon > 0$, the following bound holds

$$\Pr\left(\sup_{f \in \mathcal{H}} \left|I_S[f] - I[f]\right| \geq \epsilon\right) \leq \frac{CNM^2}{n\epsilon^2} \tag{2}$$

where $C > 0$ is some constant. What is the best C that you can get?
(Hint: use Chebychev's inequality and union bound)

(b) [**Points: 2**] Show that, if f_S is the minimizer of the empirical risk on $\mathcal{H}$, then the above inequality implies that with probability $1-\eta$ we have

$$I[f_S] \leq I_S[f_S] + \epsilon(n, \eta, N)$$

where $\epsilon(n, \eta, N) = \sqrt{\frac{CNM^2}{\eta n}}$ and $0 < \eta \leq 1$. Discuss the behavior of $I_S[f_S]$, $\epsilon(n, \eta, N)$ and their sum as functions of N.

(c) **Points: 2** Denote with f_S and f^* the minimizers on $\mathcal{H}$ of the empirical and expected risks, respectively. Given (2), show that

$$I[f_S] - I[f^*] \leq 2\epsilon(n, \eta, N).$$

(Hint: add and subtract the empirical risks of f_S and f^* in the left hand side of the above inequality. Recall that by definition f_S minimizes the empirical risk.)

Problem 5 In classification problems when the data are unbalanced (there are many more examples of one class than of the other one) a common strategy to obtain effective solution is *weighting* the loss function so that the errors in one class are counted more than errors in the other class. In the case of RLS this corresponds to solving the following problem

$$\min_{w \in \mathbb{R}^d} \left\{ \sum_{i=1}^{n} \gamma_i \big(\langle w, x_i \rangle - y_i \big)^2 + \lambda \|w\|^2 \right\}$$

where $\sum_{i=1}^{n} \gamma_i = 1$ and $\gamma_i > 0$ for all $i = 1, \ldots, n$.

(a) [**Points: 4**] Derive the explicit form of the minimizer w^* of the above problem.

(b) [**Points: 2**] Consider the case where we have a weighted loss function and we also an offset,

$$\min_{w \in \mathbb{R}^d,\, b \in \mathbb{R}} \left\{ \sum_{i=1}^{n} \gamma_i \big(\langle w, x_i \rangle + b - y_i \big)^2 + \lambda \|w\|^2 \right\}$$

where $\sum_{i=1}^{n} \gamma_i = 1$ and $\gamma_i > 0$ for all $i = 1, \ldots, n$. Using the results in PSET 1 and the above result, derive the explicit form of the minimizers w^*, b^* of the above problem.

Problem 6 In this problem we will consider a special case of sparsity based regularization and show that solution can be written in closed form. Consider the minimization problem:

$$\min_{w \in \mathbb{R}^d} \left\{ \sum_{i=1}^{n} \big(\langle w, x_i \rangle - y_i \big)^2 + 2\lambda \|w\|_1 \right\}$$

where $\langle w, x_i \rangle = \sum_{j=1}^{p} w^j x_i^j$ and $\|w\|_1 = \sum_{j=1}^{p} |w^j|$. Assume that

$$\sum_{i=1}^{n} x_i^j x_i^t = \delta_{j,t}.$$

(a) **[Points: 4]** Show that the minimizer w_* of the above problem can be written component-wise in closed form as

$$w_*^t = S_{\lambda n}(y^t)$$

where $y^t = \sum_{i=1}^n y_i x_i^t$ and

$$S_{\lambda n}(y^t) = y^t \max\{0, 1 - \frac{\lambda}{|y_t|}\}. \tag{3}$$

(Hint: Solve by computing the partial derivative of the functional w.r.t. a component w^t and setting it equal to zero. Note that you can compute the derivative of $|x|$ splitting in 3 cases: $x > 0$, $x < 0$ and $x = 0$. Also note that $\text{sign}(a) = a/|a|$.)

(b) **Points: 2** Discuss what would change if we consider

$$\min_{w \in \mathbb{R}^d,\, b \in \mathbb{R}} \left\{ \sum_{i=1}^n \big(\langle w, x_i \rangle + b - y_i\big)^2 + 2\lambda \|w\|_1 \right\}.$$

Problem 7 In this problem we will consider the problem of transductive learning, derive the explicit solution of two different learning algorithms and compare the two algorithms on a toy data set. We assume to be given a set of labeled examples $(x_i, y_i)_{i=1}^{\ell}$ and a set of unlabeled examples $(x_i)_{i=1}^{u}$, we let $m = \ell + u$. We assume that the index of the input points is ordered so that the first ℓ points have labels.
The goal is to predict the label of such an unlabeled set[1].
We consider two related schemes to do this.

We need a few concepts. We will assume a $m \times m$ (symmetric) weight matrix to be given such that similarity between two input points x_i, x_j is given by $W_{i,j}$. Then the smoothness of a function on the input points can be written as

$$R(f) = \frac{1}{2} \sum_{i,j=1}^m W_{i,j} (f(x_i) - f(x_j))^2.$$

(a) **[Points: 2]** Prove that

$$R(f) = \mathbf{f}^T L \mathbf{f} = R(\mathbf{f})$$

where $L = D - W$, D is the diagonal matrix such that $D_{ii} = \sum_{j=1}^m W_{i,j}$ and $\mathbf{f} \in \mathbb{R}^m$ with $\mathbf{f}^i = f(x_i)$, $i = 1, \ldots, m$.

(b) **[Points: 3]** Consider the algorithm

$$\min_{\mathbf{f} \in \mathbb{R}^m} R(\mathbf{f}),$$

[1] In contrast to semi-supervised learning where we might be interested into predicting labels of previously unseen points.

subject to the constraint

$$f(x_i) = y_i, \quad i = 1 \ldots, \ell.$$

Find the vector $\mathbf{f}_*$ which solves the above problem.

(c) **Points: 3** Consider the algorithm

$$\min_{\mathbf{f} \in \mathbb{R}^m} R(\mathbf{f}) + \lambda \sum_{i=1} (y_i - \mathbf{f}^i)^2.$$

Find the vector $\mathbf{f}_*^\lambda$ which solves the above problem.

(Hint for item b: write L as a block matrix with block $L_{\ell,\ell}, L_{u,\ell}, L_{\ell,u}, L_{u,u}$
Hint for item $3c$: write the output vector as $J\mathbf{y}'$ where J be the diagonal matrix defined in class which has ℓ ones on the diagonal and then zeros and $\mathbf{y}' = (y_1, \ldots, y_\ell, 0, \ldots, 0) \in \mathbb{R}^m$.

Problem 8 In this problem, we will look at gradient descent first as a regularized algorithm based on early-stopping, and then as a tool for solving RLS problems. We also recall the successive approximation scheme associated to a contractive map.

Recall that a contractive map T is such that $\|Tc - Tc'\|_2 \leq L\|c - c'\|_2$ for all c, c', with $L < 1$. By the fixed-point theorem, every contractive map has a unique fixed point: a point $c^* \in \mathbb{R}^n$ such that

$$c^* = T(c^*),$$

Then, since $\mathbb{R}^n$ is complete, it is easy to show that the iteration

$$c^{(i+1)} = T(c^{(i)}) \tag{4}$$

with $c^{(0)} = 0$, converges to the fixed point c^* for $i \to \infty$ (where superscripts denote iterates).

We have seen that the solution of ERM on a RKHS can be written as $f = \sum_{i=1}^n c_i K(\cdot, x_i)$ where the x_i belong to the training set. Then the ERM problem can be written as

$$\min_{c \in R^n} \|Y - \mathbf{K}c\|_2^2 \tag{5}$$

where $\mathbf{K}$ is the $(n \times n)$ kernel matrix and c, Y are the (n-dimensional) vectors of coefficients and labels respectively.

Let the operator-norm $\|\mathbf{K}\|$ of a matrix be the maximum absolute eigenvalue of K. By assuming that $n \geq \|\mathbf{K}\| =$ (for instance, this is always true for the Gaussian kernel), we can ensure convergence of the iterative procedures described next.

(a) **[Points: 2]** Prove that c^* minimizes the empirical risk in (5), if and only if it satisfies $c^* = T(c^*)$ with

$$T(c) := c - \frac{1}{n}(\mathbf{K}c - Y). \tag{6}$$

(b) [**Points: 2**] Now recall the Tikhonov regularization problem

$$\min_{c\in\mathbb{R}^n}\{\|Y-\mathbf{K}c\|_2^2+\lambda c^t\mathbf{K}c\}. \tag{7}$$

Show that c_λ^* solves the problem (7) if and only if it also satisfies $c_\lambda^* = T(c_\lambda^*)$ with

$$T_\lambda(c) := c - \frac{1}{n+\lambda}((\mathbf{K}+\lambda I)c - Y). \tag{8}$$