MIT 9.520/6.860: Statistical Learning Theory and Applications **Fall 2018**

Problem Set 1

Due Date: **Sat., Sep 29 2018, 11:59 pm** *(Stellar submission)*

Instructions: There are **4 problems** in total in this problem set. The breakdown of individual scores per sub-problem are provided. Use the provided LaTeX template to typeset your report. Provide sufficient explanations in all solutions but avoid proving lecture or out-of-scope material (unless explicitly asked to). An 8-page submission maximum is allowed (do not change the font or margin of the template).

What to submit: Submit your report **online through Stellar** by the due date/time. Submission must be a single pdf in LaTeX format. Include code files separately, if applicable. Submit a **printout of your report** in the first class after the due date.

Policies: Collaborative reports are not allowed. Even if you discuss problems with classmates, you are expected to write and submit **individual reports**.

Problem 1 [25 points] In (binary) classification problems the classification or "decision" rule is a binary valued function $c : X \to Y$, where $Y = \{1, -1\}$. The quality of a classification rule can be measured by the misclassification error

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

If we introduce the misclassification loss $\Theta(-yc(x))$, where $\Theta(\alpha) = 1$ if $\alpha > 0$ and $\Theta(\alpha) = 0$ otherwise, the misclassification error can be rewritten as

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

In practice, one usually looks for real valued functions $f : X \to \mathbb{R}$ and replaces $\Theta(-yc(x))$ with some convex loss $\ell(-yf(x))$, with $\ell : \mathbb{R} \to [0, \infty)$. A classification rule is obtained by taking $c(x) = \text{sign}(f(x))$, and the error is measured by the expected error

$$L(f) = \int_{X \times Y} \ell(-yf(x))p(x)p(y|x)dxdy.$$

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

With the above discussion in mind, and *assuming that the distribution $p(x, y)$ is known*:

1.1 Derive the explicit form of the minimizer of $L(f)$ for the:

a) exponential loss $\ell(-yf(x)) = \exp(-yf(x))$,

b) logistic loss $\ell(-yf(x)) = \log(1 + \exp(-yf(x)))$.

c) hinge loss $\ell(-yf(x)) = |1 - yf(x)|_+$.

d) the misclassification loss $\Theta(-yc(x))$.

1.2 Discuss how the target functions of the exponential, logistic and hinge loss functions relate to the target function of the misclassification loss.

Problem 2 [30 points] You are going to the derive an alternative proof of the representer theorem that holds very generally but does not give an explicit expression for the obtained coefficients. Then you are going to compare the gradient descent solution for logistic regression using this result to the one derived in class.

2.1 Consider regularized least squares

$$\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} (y_i - w^\top x_i)^2 + \lambda \|w\|^2.$$

Show that the solution of the above problem is of the form $\hat{w} = \sum_{i=1}^n x_i c_i$. But to do this, start from the observation that any $w \in \mathbb{R}^d$ can be written as $w = w_n + w_n^\perp$, where w_n is of the desired form, i.e. $\hat{w} = \sum_{i=1}^n x_i c_i$, and $w_n^\top w_n^\perp = 0$.

2.2 Show that the above proof generalizes to problems of the form

$$\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, w^\top x_i) + \lambda \|w\|^2$$

where ℓ is convex.

2.3 Using the above result we can now consider the problem

$$\min_{c \in \mathbb{R}^n} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, \sum_{j=1}^{n} x_j^\top x_i c_i) + \lambda \sum_{j=1}^{n} \sum_{i=1}^{n} x_j^\top x_i c_j c_i.$$

For the logistic loss, start from this latter expression, and derive a corresponding gradient descent iteration. Compare to the iteration derived in class.

Problem 3 [15 points] A 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) offset term in the case of linear functions.

Consider Tikhonov regularization in the linear case, but assume that there is an unpenalized offset term b,

$$\min_{w\in\mathbb{R}^d,\, b\in\mathbb{R}} \left\{ \frac{1}{n}\sum_{i=1}^{n} \big(\langle w, x_i\rangle + b - y_i\big)^2 + \lambda\|w\|^2 \right\}, \tag{1.1}$$

and let (w^*, b^*) be the solution of the above problem. Denote by $x_i^c = x_i - \bar{x}$, $y_i^c = y_i - \bar{y}$ the centered data for $i = 1, \ldots, n$, where $\bar{y}, \bar{x}$ are the output and input means respectively.

3.1 Show that w^* also solves

$$\min_{w\in\mathbb{R}^d} \left\{ \frac{1}{n}\sum_{i=1}^{n} \big(\langle w, x_i^c\rangle - y_i^c)^2 + \lambda\|w\|^2 \right\}. \tag{1.2}$$

3.2 Give closed form expressions for the solutions $b^\star$ and $w^\star$.

Problem 4 [20 points] The distance between two elements $\Phi(x), \Phi(x')$ of a feature space $\mathcal{F}$ induced by some kernel K can be seen as a new distance $d_K(x, x')$ in the input X.

4.1 Show that such a distance can always be calculated without knowing the explicit form of the feature map Φ itself.

4.2 Consider a dataset of pairs $\{(x_i, y_i)\}_{i=1}^{N}$, with $x_i \in X$ and $y_i \in \{-1, 1\}$, such that n_+ of the x_i have label $+1$ and n_- have label -1 ($n_+ + n_- = N$). 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$(\cdot)$ function), that assigns to a test point x the label of the class whose mean is closest *in the feature space* according to the distance d_K.