MIT 9.520/6.860: Statistical Learning Theory and Applications **Fall 2022**

Problem Set 1

Due Date: **Tue., Oct 4 2022, 10:59 am** *(Canvas submission)*

Instructions: There are **5 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).

What to submit: Submit your report **online through Canvas** by the due date/time. Submission must be a single pdf in LaTeX format.

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

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

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 [20 points] Derive the explicit form of the minimizer of $\mathcal{E}(f)$ 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 [5 points] State how the target functions of the exponential, logistic and hinge loss functions are related to the target function of the misclassification loss.

Problem 2 [20 points] In this problem we will derive an alternative proof of the representer theorem that holds very generally but does not give an explicit expression for the obtained coefficients. Then we will compare the gradient descent solution for logistic regression using this result to the one derived in class.

2.1 [10 points] 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 [5 points] 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 [5 points] Using the above result we can now consider the problem

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

For the logistic loss, start from this latter expression, and derive a corresponding gradient descent iteration. Compare its computational complexity (in time, and in memory) to the computational complexity of the w gradient descent iteration seen in class and reproduced below:

$$w_{t+1} = w_t - \eta \left(-\frac{1}{n} \sum_{i=1}^{n} \frac{x_i y_i}{1 + e^{y_i w_t^T x_i}} + 2\lambda w_t \right)$$

Problem 3 [15 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.

3.1 [5 points] Show that such a distance can always be calculated without knowing the explicit form of the feature map Φ itself (that is, in terms of $K(x, x')$).

3.2 [10 points] 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.

Problem 4 [20 points] In this problem, we will study the gradient descent iteration to understand aspects of convergence. First we recall the successive approximation scheme associated to a contractive map.
Recall that a contractive map T is such that $\|T(w) - T(w')\|_2 \leq L\|w - w'\|_2$ for all $w, w' \in \mathbb{R}^d$, with $L < 1$. By the Banach fixed-point theorem every contractive map has a unique fixed point: a point $w^* \in \mathbb{R}^d$ such that $w^* = T(w^*)$. Then, since $\mathbb{R}^d$ is complete, it is easy to show that the iteration $w^{(i+1)} = T(w^{(i)})$ with $w^{(0)} = 0$, converges to the fixed point w^* for $i \to \infty$ (where superscripts denote iterates).
Consider the empirical risk minimization problem

$$\min_{w \in \mathbb{R}^d} \frac{1}{2n} \|Y - Xw\|_2^2 \tag{1.1}$$

where $x_i \in X = \mathbb{R}^d$, $X = (x_1, \ldots, x_n)^\top$, is the $(n \times d)$ data matrix whose rows correspond to the training points and Y the $(n \times 1)$ vector of labels.

Let the operator-norm $\|X\|$ of a matrix X to be the maximum singular value of X. Assume the input vectors to have all norm $\|x\| \leq 1$ and that the matrix X has rank d.

4.1 [10 points] Prove that w^* minimizes the empirical risk in (1.1), if and only if it satisfies $w^* = T(w^*)$ with

$$T(w) := w - \frac{1}{n} X^\top (Xw - Y). \tag{1.2}$$

Grad_w(L) = 0 and T(w)=x give the same set of equations!!!

Then, prove whether T is a contraction.

4.2 [10 points] For the ERM problem (1.1) the gradient descent iteration is given by

$$w^{(i+1)} = w^{(i)} - \gamma \frac{1}{n} X^\top (Xw^{(i)} - Y). \tag{1.3}$$

where $\gamma > 0$ is a step-size.
Classical results show that a suitable choice for the step-size is $\gamma = 2/L$, where L is the Lipschitz constant of the gradient of the objective function. Compute L (hence choose a corresponding step-size) explicitly. Also, derive an explicit estimate of L for the case in which $\|x_i\| \leq 1$, for all $i = 1, \ldots, n$.

$tr(X^\top X) = tr(XX^\top)$

Problem 5 [20 points] In learning from imbalanced data (i.e., there are many more examples of one class than of the other) a common strategy is to *weight* the loss function so that the errors in one class are counted more than those of the other class. In the case of linear RLS, this corresponds to solving the modified 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 the weights $\{\gamma_i : i = 1, \ldots, n\}$ are chosen so that $\sum_{i=1}^{n} \gamma_i = 1$ and $\gamma_i > 0$ for all $i = 1, \ldots, n$.

5.1 Derive the explicit form of the minimizer w^*.

5.2 Consider the case of a weighted loss function *and* an unpenalized offset $b \in \mathbb{R}$,

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

Derive the explicit form of the minimizers w^*, b^* for this problem.