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

Problem Set 1

Student Name: Sebastiani Aguirre-Navarro

Problem 1 1.1 Fixing $f(x)$ in the loss function and then minimizing the inner integral with respect to this fixed value can give us the target function.

1. Exponential Loss

   Let $l(-yf(x)) = exp(-yf(x))$, then then the target function that minimizes:

   $$\min_f \int_{y\sim p(y|x)} exp(-yf)p(y|x)dy$$

   Then

   $$0 = \frac{\partial}{\partial f}\left[\int_{y\sim p(y|x)} exp(-yf)p(y|x)dy\right]$$

   Then

   $$0 = \int_{y\sim p(y|x)} -yexp(-yf)p(y|x)dy$$

   However, $y \in \{-1, +1\}$

   $$0 = exp(f)p(-1|x) - exp(-f)p(1|x)$$

   And $p(-1|x) = 1 - p(1|x)$

   $$0 = exp(f)(1 - p(y|x)) - exp(-f)p(1|x)$$

   $$\begin{aligned} exp(f)(1 - p(1|x)) &= exp(-f)p(1|x) \\ \frac{exp(f)}{exp(-f)} &= \frac{p(1|x)}{1 - p(1|x)} \\ ln(exp(2f)) &= ln\left(\frac{p(1|x)}{1 - p(1|x)}\right) \end{aligned} \tag{1.1}$$

   The target function is then

   $$f^* = \frac{1}{2} ln\left(\frac{p(1|x)}{1 - p(1|x)}\right)$$

2. Logistic Loss

Let $l(-yf(x)) = log(1 + exp(-yf(x)))$ Then

$$\begin{aligned}
&\min_f \int_{y\sim p(y|x)} log(1+exp(-yf))p(y|x)dy \\
&0 = \frac{\partial}{\partial f}\int_{y\sim p(y|x)} log(1+exp(-yf))p(y|x)dy \\
&0 = \int_{y\sim p(y|x)} \frac{-yexp(-yf)}{1+exp(-yf)}p(y|x)dy \\
&0 = \frac{exp(f)}{1+exp(f)}p(-1|x) - \frac{exp(-f)}{1+exp(-f)}p(1|x) \qquad (1.2)\\
&0 = \frac{1}{1+exp(-f)}(1-p(1|x)) - \frac{1}{1+exp(f)}p(1|x) \\
&\frac{p(1|x)}{1-p(1|x)} = \frac{1+exp(f)}{1+exp(-f)} \\
&ln\left(\frac{p(1|x)}{1-p(1|x)}\right) = ln(1+exp(f)) - ln(1+exp(f)) + ln(exp(f))
\end{aligned}$$

The target function is then

$$f^* = ln\left(\frac{p(1|x)}{1-p(1|x)}\right)$$

3. Hinge Loss

   Let $l(-yf(x)) = |1 - yf(x)|$

$$\begin{aligned}
&\min_f \int_{y\sim p(1|x)} |1-yf|\, p(y|x)dy \\
&= \min_f |1-f|_+ p(1|x) + |1+f|_+ p(-1|x) \qquad (1.3)\\
&= \min_f [max\{1-f,0\}p(1|x) + max\{1+f,0\}p(-1|x)]
\end{aligned}$$

   If $p(1|x) > p(-1|x)$, then $f \geq 1$ such that the term $p(1|x)$ vanishes and only $p(-1|x)$ remains.

   If $p(-1|x) > p(1|x)$, then $f \leq -1$ such that the term $p(-1|x)$ vanishes and only $p(+1|x)$ remains.

   Therefore, $f^* = p(1|x) - p(-1|x)$

4. Missclassification Loss

   In this case, we now try to find the function $c(x) : X \mapsto \{-1, +1\}$. Let now the loss function be

$$\theta(-yc(x)) = \begin{cases} 1 & -yc(x) < 0 \\ 0 & -yc(x) > 0 \end{cases} \qquad (1.4)$$

Then the minimization problem is

$$\begin{aligned}&\min_f \int_{y\sim p(y|x)} \theta(-yc(x))p(y|x)dy \\ &= \min_f \theta(-f)p(1|x) + \theta(f)p(-1|x)\end{aligned} \tag{1.5}$$

following the same argument as above,

if $p(1|x) > p(-1|x)$, then $f^* > 0$

if $p(1|x) < p(-1|x)$, then $f^* < 0$

then it follows that $c(x) = sign(p(1|x) - p(-1|x))$

1.2

The relationship between the target functions of the surrogate loss functions and the target function for the missclassification loss relate in that if you take the sign function from the surrogate ones, you can convert them into the same nature as the missclassification target function. In the case of the Hinge Loss, it gives you the same relationship between the class conditional probabilities as the missclassification loss.

Problem 2 2.1 Let the solution be expressed as $w = w_n + w_n^\perp$(the sum of a particular solution and a null solution), with $w_n w_n^\perp = 0$ an that $w_n = \sum_i^d c_i x_i$. It follows from these supossisions that

$$w_n w_n^\perp = \sum_i^d c_i x_i^T w_n^\perp = 0$$

which implies that $x_i \perp w_n^\perp, \forall x_i \in (x_i, y_i)_{i=1}^N$. Now consider let us consider the empirical risk minimization problem

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

Then

$$\begin{aligned}&\min_{w\in\mathcal{R}} \frac{1}{n}\sum_{i=1}^N (y_i - (w_n + w_n^\perp)^T x_i)^2 + \lambda||w_n + w_n^\perp||^2 \\ =&\min_{w\in\mathcal{R}} \frac{1}{n}\sum_{i=1}^N (y_i - w_n^T x_i + w_n^{\perp,T} x_i)^2 + \lambda||w_n + w_n^\perp||^2 \\ =&\min_{w\in\mathcal{R}} \frac{1}{n}\sum_{i=1}^N (y_i - w_n^T x_i)^2 + \lambda||w_n + w_n^\perp||^2\end{aligned} \tag{1.6}$$

That last step is by using the fact that $x_i \perp w_n^\perp$. The norm reduces to

$$\lambda(w_n^T + w_n^{\perp,T})(w_n + w_n^\perp) = \lambda(w_n^T w_n + 2w_n^T w_n^\perp + w_n^{\perp,T} w_n^\perp)$$

Since $w_n w_n^\perp = 0$, this reduces to $\lambda||w_n||^2 + \lambda||w_n^\perp||^2$. Since the empirical risk does not depend directly on $w_n^\perp$ and $\lambda > 0$, then minimization occurrs if $||w_n^\perp||^2 = 0$. Because w_n is the solution of this minimization problem, it follows that $w = \sum_{i=1}^N c_i x_i$.

2.2

Let the suppositions be the same as before. Let us consider the risk minimization problem

$$\min_{w\in\mathcal{R}^d} \frac{1}{n}\sum_{i=1}^{N} l(y_i, w^T x_i) + \lambda||w||^2$$

Where $l(y_i, w^T x_i)$ is any convex loss function. We can express this ERM as:

$$\min_{w\in\mathcal{R}^d} \frac{1}{n}\sum_{i=1}^{N} l(y_i, (w_n + w_n^\perp)^T x_i) + \lambda||w_n||^2 + \lambda||w_n^\perp||^2$$

However, as proved before, $x_i \perp w_n^\perp, \forall x_i \in \{(x_i, y_i)\}_{i=1}^N$, then

$$\min_{w\in\mathcal{R}^d} \frac{1}{n}\sum_{i=1}^{N} l(y_i, w_n^T x_i) + \lambda||w_n||^2 + \lambda||w_n^\perp||^2$$

Again, this implies that $\lambda||w_n^\perp||^2 = 0$, since $\lambda > 0$. Therefore, since the problem depends explicitly on w_n, the w that solves this problem can be written in the form $w = \sum_{i=1}^N c_i x_i$.

2.3

Let the loss function be the logistic function. Using $w = \sum_{i=1}^N c_i x_i$, the risk minimization problem is

$$\begin{aligned}
&\min_{c_j} \frac{1}{n}\sum_{i=1}^{N} log(1 + exp(-y_i \sum_{j=1}^{N} x_j^T x_i c_j)) + \lambda \sum_{j=1}^{N}\sum_{i=1}^{N} x_j^T x_i c_j c_i \\
=&\frac{\partial}{\partial c_j}\left[\frac{1}{n}\sum_{i=1}^{N} log(1 + exp(-y_i \sum_{j=1}^{N} x_j^T x_i c_j)) + \lambda \sum_{j=1}^{N}\sum_{i=1}^{N} x_j^T x_i c_j c_i\right] \\
=&\frac{1}{n}\sum_{i=1}^{N} \frac{-y_i x_i^T x_j exp(-y_i \sum_{j=1}^{N} x_j^T x_i c_j)}{1 + exp(-y_i \sum_{j=1}^{N} x_j x_i c_j)} + \lambda \sum_{j=1}^{N}\sum_{i=1}^{N} x_j^T x_i c_i \\
=&\frac{1}{n}\sum_{i=1}^{n} \frac{-y_i x_j^T x_j}{1 + exp(y_i \sum_{i}^{n} x_j^T x_i c_j)} + \lambda \sum_{i}^{N} x_j^T x_i c_i \\
=&\frac{1}{n}\sum_{i}^{n} \frac{-y_i x_j^T x_i}{1 + exp(y_i \sum_{i}^{n} x_j^T x_i c_j)} + \lambda x_j^T x_i c_i \\
=&\frac{1}{n} x^T \left[\sum_{i}^{n} \left(\frac{-y_i}{1 + exp(y_i \sum_{j}^{n} x_j x_i c_j)} + \lambda c_i\right) x_i\right]
\end{aligned} \tag{1.7}$$

This last expression is the gradient of the empirical risk plus regularization, then the rule to update each c_j is

$$c_j \leftarrow c_j - \gamma\left(\frac{\partial}{\partial c_j} erm(w) + ||w||^2\right)$$

Replacing w accordingly with the representer theorem. If you take the gradient of the version discussed in class and express the w in a linear combination of the input data, you get the gradient needed to update the c_j.

Problem 3 3.1

Let w^* be the solution that minimizes the loss function with a constant bias b as described in the problem. Let us consider:

$$L^c(w) = \frac{1}{n}\sum_{i=1}^{n}(\langle w, x_c^i\rangle - y_c^i)^2 + \lambda||w||^2$$

Where $x_c = x_i - \bar{x}$ and $y_c = y_i - \bar{y}$. Evaluating $L^c(w)$ at w^*, we obtain

$$\begin{aligned} L^c(w^*) &= \frac{1}{n}\sum_{i=1}^{n}(\langle w^*, x_i - \bar{x}\rangle - y_i - \bar{y})^2 + \lambda||w||^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}(\langle w^*, x_i\rangle - \langle w^*, \bar{x}\rangle - y_i + \bar{y})^2 + \lambda||w||^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}(\langle w^*, x_i\rangle + (\bar{y} - \langle w^*, \bar{x}\rangle) - y_i)^2 + \lambda||w||^2 \end{aligned} \tag{1.8}$$

because w^* is a fixed point, and $\bar{y}$, $\bar{x}$ are constants, then $\bar{y} - \langle w^*, \bar{x}\rangle$ is also constant, then the above expression is equal to

$$L(w) = \frac{1}{n}\sum_{i=1}^{n}(\langle w, x_i\rangle + b - y_i)^2 + \lambda||w||^2$$

Which is the original loss function that w^* solves. Therefore, w^* solves $L^c(w)$.

3.2

Now let us find the parameters that minimize the empirical risk:

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

Then, differentiating with respect to b

$$\begin{aligned} 0 &= \frac{\partial}{\partial b}\left[\frac{1}{n}\sum_{i=1}^{n}(\langle w, x_i\rangle + b - y_i)^2 + \lambda||w||^2\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}2(\langle w, x_i\rangle + b^* - y_i) \\ &= \frac{1}{n}\left[\sum_{i=1}^{n}(\langle w, x_i\rangle - y_i) + nb^*\right] \\ b^* &= \bar{y} - \langle w, \bar{x}\rangle \end{aligned} \tag{1.9}$$

Differentiating with respect to w

$$\begin{aligned} 0 &= \frac{\partial}{\partial w}\left[\frac{1}{n}\sum_{i=1}^{n}(\langle w, x_i\rangle + b - y_i)^2 + \lambda||w||^2\right] \\ &= \frac{2}{n}\left[\sum_{i=1}^{n}(\langle w^*, x_i\rangle + b^* - y_i)x_i\right] + 2\lambda w^* \\ &= \frac{1}{n}\sum_{i=1}^{n}\langle w^*, x_i\rangle x_i + \frac{b}{n}\sum_{i=1}^{n}x_i - \frac{1}{n}\sum_{i=1}^{n}y_i x_i + \lambda w^* \end{aligned} \tag{1.10}$$

$$\frac{1}{n}\sum_{i=1}^{n}x_i y_i - b^*\bar{x} = \frac{1}{n}\sum_{i=1}^{n}\langle w^*, x_i\rangle x_i + \lambda w^*$$

$$w^* = (\overline{xx^T} - \bar{x}\bar{x}^T + \lambda I_{d\times d})^{-1}(\overline{xy} - \bar{y}\bar{x})$$

Since $\overline{xx^T} - \bar{x}\bar{x}^T$ and $\overline{xy} - \bar{y}\bar{x}$ are just covariances, then

$$w^* = (X^TX + \lambda I_{d\times d})^{-1}X^TY$$

Problem 4 4.1

Let the distance between two points in $\mathcal{F}$ be

$$d_k(x, x') = ||\Phi(x) - \Phi(x')||^2$$

then

$$\begin{aligned} &= (\Phi(x) - \Phi(x'))^T(\Phi(x) - \Phi(x')) \\ &= \Phi(x)^T\Phi(x) + \Phi(x')^T\Phi(x') - 2\Phi(x)\Phi(x') \end{aligned} \tag{1.11}$$

Since $\mathcal{F}$ is induced by a kernel, $K(x, x') = \langle\Phi(x), \Phi(x')\rangle$, then

$$d_k(x, x') = K(x, x) + K(x', x') - 2K(x, x')$$

Which means that to calculate the distance of two points in feature space, the explicit form of the feature map is not needed since it is in terms of the kernel.

4.2

Let $(x_i, y_i)_{i=1}^N$, where n_+ is the number of x_i that have label $y_i = +1$ and n_- is the number of x_i that have label $y_i = -1$. Then the mean of the two classes are

$$\mu_{+1} = \frac{1}{n_+}\sum_{i=1}^{n_+}x_i$$

$$\mu_{-1} = \frac{1}{n_-}\sum_{i=1}^{n_-}x_i$$

Using the definition of distance from the previous subproblem, then the rule would be

$$c(x) = sign(d_K(x, \mu_{-1}) - d_K(x, \mu_{+1}))$$

This rule assigns a test point x the class whose mean is closest in feature space. If the test point is closest to class -1, then the subtraction is negative, and if it is closest to $+1$ then the subtraction is positive, thus assigning the appropriate labels.