MIT 9.520: Statistical Learning Theory **Fall 2014**

Problem Set 1

Student: Brando Miranda

Problem 1

Problem 2 Please write your analysis on Problem 2 here

Problem 3 a) We want to show that $d(\Phi(x), \Phi(x')) = d_k(x, x')$.

$$d(\Phi(x), \Phi(x')) = ||\Phi(x) - \Phi(x')|| = \sqrt{\langle \Phi(x) - \Phi(x'), \Phi(x) - \Phi(x') \rangle}$$

by linearity and symmetry of the inner product:

$$\sqrt{\langle \Phi(x), \Phi(x) \rangle - 2\langle \Phi(x), \Phi(x') \rangle + \langle \Phi(x'), \Phi(x') \rangle}$$

by the reproducing property of the kernel K we know $K(x, x) = \langle K_x, K_x \rangle = \langle \Phi(x), \Phi(x) \rangle$. Thus:

$$\sqrt{K(x, x) + K(x', x') - 2K(x, x')}$$

Thus:

$$d(\Phi(x), \Phi(x')) = d_k(x, x') = \sqrt{K(x, x) + K(x', x') - 2K(x, x')}$$

Which is only a function of the input vector x and does not need the explicit representation of the feature map $\Phi(x)$

b) Let x_+ denote the n_+ data points with positive label $y = +1$ and similar for x_- for points with negative labels.

$$d(\Phi(x), \sum_{x_+ s.t. y_i = +1} \Phi(x_+)) = \sqrt{\langle \Phi(x'), \sum_{x_+ s.t. y = +1} \Phi(x_+) \rangle}$$

by linearity and symmetry of inner product we get:

$$\sqrt{\sum_{x_+ s.t. y = +1} \langle \Phi(x'), \Phi(x_+) \rangle}$$

Problem 4 a) To check that the square loss function can be written as $\mathcal{L}(-yf(x))$ lets expand $\|f(x) - y\|^2$:

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

but $y^2 = 1$ thus:

$$\mathcal{L}(-yf(x)) = (1 - 2yf(x) + (yf(x))^2)$$

To find the minimizer c(x) we need to minimize:

$$\mathbb{E}_{x,y}[(y - f(x))^2]$$

and specify the function that achieves this minimum. Lets find it by taking the derivative of the above wrt to f(x) and setting it to zero:

$$\frac{d}{df(x)}\mathbb{E}_x\mathbb{E}_{y|x}[(y - f(x))^2] = \mathbb{E}_x\frac{d}{df(x)}\mathbb{E}_{y|x}[(y - f(x))^2]$$

which can be minimized by finding the minimum of $\frac{d}{df(x)}\mathbb{E}_{y|x}[(y - f(x))^2]$:

$$\frac{d}{df(x)}\mathbb{E}_{y|x}[(y - f(x))^2] = \mathbb{E}_{y|x}\left[\frac{d}{df(x)}(y - f(x))^2\right] = 0$$

$$\mathbb{E}_{y|x}[2(y - f(x))] = 0$$

$$\mathbb{E}_{y|x}[y] = \mathbb{E}_{y|x}[f(x)]$$

$$\mathbb{E}_{y|x}[y] = f(x)\mathbb{E}_{y|x}[1]$$

$$\mathbb{E}_{y|x}[y] = f(x)$$

$$p_{y|x}(1|x) - p_{y|x}(-1|x) = f(x)$$

Since $p_{y|x}(1|x) + p_{y|x}(-1|x) = 1$ then:

$$2p_{y|x}(1|x) - 1 = f(x)$$

b) We want to solve:

$$f^*(x) = argmin_{f(x)}\mathbb{E}_{x,y}[e^{-yf(x)}]$$

$$\frac{d}{df(x)}\mathbb{E}_x\mathbb{E}_{y|x}[e^{-yf(x)}] = 0$$

Similar reasoning as the previous question we have:

$$\mathbb{E}_{y|x}\left[\frac{d}{df(x)}e^{-yf(x)}\right] = 0$$

$$\sum_{y\in\{1,-1\}} p_{y|x}(y|x)ye^{-yf(x)} = p_{y|x}(1|x)e^{-f(x)} - p_{y|x}(-1|x)e^{f(x)}$$

$$p_{y|x}(1|x) - p_{y|x}(-1|x)e^{2f(x)} = 0$$

$$p_{y|x}(1|x) = p_{y|x}(-1|x)e^{2f(x)}$$

$$\frac{p_{y|x}(1|x)}{p_{y|x}(-1|x)} = e^{2f(x)}$$

$$\frac{p_{y|x}(1|x)}{p_{y|x}(-1|x)} = e^{2f(x)}$$

$$\frac{1}{2}log\left(\frac{p_{y|x}(1|x)}{p_{y|x}(-1|x)}\right) = f(x)$$

or

$$\frac{1}{2}log\left(\frac{p_{y|x}(1|x)}{1-p_{y|x}(1|x)}\right) = f(x)$$

c) When we apply a function that is monotonic to another function, then the value that minimizes it does not change. Said differently, if we have a function that preserves monotonicity (and thus preserves order), then the minimizer does not change. i.e. if $f(x) < f(y)$ and $g(x)$ is monotonic then $g(f(x)) < g(f(y))$ and because of that the value of x that minimized $f(x)$ also minimizes $g(f(x))$.

The function $g(x) = x + 1$ is clearly monotonic. So is the function $h(x) = log(x)$. Now consider the following function:

$$h(g(e^{-yf(x)})) = log(g(e^{-yf(x)})) = log(1 + e^{-yf(x)})$$

This time we are trying to minimize:

$$\mathcal{L}(-yf(x)) = (h(g(e^{-yf(x)}))$$

From part b) we notice that its just a composite function of the exponential loss function using two monotonic functions. So without the need of further calculations its clear that the minimizer is the same as part b:

$$\frac{1}{2}log\left(\frac{p_{y|x}(1|x)}{p_{y|x}(-1|x)}\right) = f(x)$$

or

$$\frac{1}{2}log\left(\frac{p_{y|x}(1|x)}{1-p_{y|x}(1|x)}\right) = f(x)$$

d) Bayes decision rule is:

$$b(x) = sign(2p_{y|x}(1|x) - 1)$$

From part a, b and c we have $f(x)$ expressed in terms of $p_{y|x}(1|x)$. Thus, we can just re-arrange those equations and make $p_{y|x}(1|x)$ the subject and therefore, express $p_{y|x}(1|x)$ as a function of $f(x)$. Then we can obviously plug them back to $b(x)$ that is a function of $p_{y|x}(1|x)$. To not bore you with the simple algebra I will just express the answers.

For part a) we have:

$$p_{y|x}(1|x) = \frac{f(x)+1}{2}$$

So the relation of the minimizer of the squared loss function to Bayes decision rule is:

$$b(x) = sign(2\left(\frac{f(x)+1}{2}\right) - 1) = sign(f(x))$$

Since b) and c) have the same minimizer, then their relation to Bayes decision rule is the same. First lets express $p_{y|x}(1|x)$ interns of f(x):

$$p(1|x) = \frac{1}{1+e^{-f(x)}}$$

$$b(x) = sign\left(\frac{2}{1+e^{-f(x)}} - 1\right)$$

Problem 5 (MATLAB) Please write your analysis on Problem 5 here