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

Problem Set 1

Student: Brando Miranda

Problem 1 The optimal solution to the Tikhonov regularization with offset (but not penalized) and linear kernel must satisfy the following constraints for minimization/optimality:

$$\frac{d}{db}\frac{1}{n}\sum_{i=1}^{n}(\langle w, x_i\rangle_{\mathbb{R}^d} + b - y_i)^2 + \lambda||w||^2_{\mathbb{R}^d} = 0$$

$$b = \frac{1}{n}\sum_{i=1}^{n}(y_i - \langle w, x_i\rangle_{\mathbb{R}^d}) = \frac{1}{n}\sum_{i=1}^{n} y_i - \langle w, \frac{1}{n}\sum_{i=1}^{n} x_i\rangle_{\mathbb{R}^d}$$

let $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ then b:

$$b = \bar{y} - \langle w, \bar{x}\rangle_{\mathbb{R}^d}$$

If the above must be satisfy, then we can just "remove" the dependence of the offset b, of the original minimization by substituting the above value of b. That leads to the following:

$$min_{w\in\mathbb{R}^d}\{\frac{1}{n}(\langle w, x_i\rangle_{\mathbb{R}^d} + \bar{y} - \langle w, \bar{x}\rangle_{\mathbb{R}^d} - y_i)^2 + \lambda||w||^2_{\mathbb{R}^d})\}$$

by using linearity of the inner product and re-arranging the y terms, we get the following expression:

$$min_{w\in\mathbb{R}^d}\{\frac{1}{n}(\langle w, x_i - \bar{x}\rangle_{\mathbb{R}^d} - (y_i - \bar{y}))^2 + \lambda||w||^2_{\mathbb{R}^d})\}$$

$$min_{w\in\mathbb{R}^d}\{\frac{1}{n}(\langle w, x_i^c\rangle_{\mathbb{R}^d} - y_i^c)^2 + \lambda||w||^2_{\mathbb{R}^d})\}$$

Which if we just substitute the definitions of x_i^c and y_i^c, we notice that it is the same minimization problem as with centered data! Thus, the solution for w^* for the original problem with unpenalized offset, is the same as when the data is centered. QED.

Let X^c be defined as the data matrix but with the centered data as the rows, instead of the input data. Similarly define Y^c. Then we can express w^* in terms of X^c and Y^c

$$w^* = ((X^c)^T X^c + \lambda n I)^{-1}(X^c)^T Y^c$$

Which we can then plug back in to out original formulation for b from the beginning yields:

$$b = \bar{y} - \langle w^*, \bar{x}\rangle_{\mathbb{R}^d}$$

We can do this because the optimal value for w^* must also satisfy that the derivative wrt b is zero (which is the same as the above equation for b).

$$b = \bar{y} - \langle((X^c)^T X^c + \lambda n I)^{-1}(X^c)^T Y^c, \bar{x}\rangle_{\mathbb{R}^d}$$

$$b = \bar{y} - ((X^c)^T X^c + \lambda n I)^{-1}(X^c)^T Y^c)^T \bar{x}$$

Problem 2 Lets try to express the minimization problem in terms of matrices:

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

Let

$$\hat{\gamma} = \begin{pmatrix} \gamma_1 & 0 & 0 & 0 \\ 0 & \gamma_2 & 0 & 0 \\ 0 & 0 & \ddots & \vdots \\ 0 & 0 & \cdots & \gamma_n \end{pmatrix}$$

and lets try to re-write the minimization problem in terms of the above gamma matrix:

$$(\hat{\gamma}(Xw - Y)) \cdot (Xw - Y) + \lambda||w||^2_{\mathbb{R}^d}$$

$$(\hat{\gamma}(Xw - Y))^T(Xw - Y) + \lambda||w||^2_{\mathbb{R}^d}$$

$$\langle\hat{\gamma}(Xw - Y), Xw - Y\rangle_{\mathbb{R}^n} + \lambda||w||^2_{\mathbb{R}^d}$$

Lets try to minimize it now by taking the gradient wrt w:

$$\bigtriangledown_w \left(\langle\hat{\gamma}(Xw - Y), Xw - Y\rangle_{\mathbb{R}^n} + \lambda||w||^2_{\mathbb{R}^d}\right) = 0$$

Lets handle the inner product first:

$$\bigtriangledown_w \left(\langle\hat{\gamma}(Xw - Y), Xw - Y\rangle_{\mathbb{R}^n}\right)$$

By applying linearity of the inner product we get:

$$\langle\hat{\gamma}Xw - \hat{\gamma}Y, Xw - Y\rangle_{\mathbb{R}^n} = \langle\hat{\gamma}Xw, Xw\rangle_{\mathbb{R}^n} - \langle\hat{\gamma}Xw, Y\rangle_{\mathbb{R}^n} - \langle\hat{\gamma}Y, Xw\rangle_{\mathbb{R}^n} + \langle\hat{\gamma}Y, Y\rangle_{\mathbb{R}^n}$$

$$\bigtriangledown_w((\hat{\gamma}Xw)^T Xw - (\hat{\gamma}Xw)^T Y - Y^T\hat{\gamma}Xw + (\hat{\gamma}Y)^T Y + \lambda n||w||^2_{\mathbb{R}^d}) = 0$$

$$2X^T\hat{\gamma}Xw - X^T\hat{\gamma}Y - Y^T\hat{\gamma}X + 2\lambda nw = 0$$

$$X^T\hat{\gamma}Xw + \lambda nw = Y^T\hat{\gamma}X$$

$$(X^T\hat{\gamma}X + \lambda nI)w = Y^T\hat{\gamma}X$$

$$w = (X^T\hat{\gamma}X + \lambda nI)^{-1}Y^T\hat{\gamma}X$$

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 $K(x, x) = \langle K_x, K_x \rangle = \langle \Phi(x), \Phi(x) \rangle$:

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

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 positively label $y = +1$, X_+ the set containing them and $\mu_{I_+} = \frac{1}{n_+} \sum_{x_+ \in X_+} \Phi(x)$. Similarly, define x_-, X_- and μ_{I_-}. Let the average distance in the feature space be:

$$d(\Phi(x), \mu_{I_+}) = \sqrt{\langle \Phi(x) - \frac{1}{n_+} \sum_{x_+ \in X_+} \Phi(x_+), \Phi(x) - \frac{1}{n_+} \sum_{x_+ \in X_+} \Phi(x_+) \rangle}$$

by linearity and symmetry of inner products we get:

$$\sqrt{\langle \Phi(x), \Phi(x) \rangle - \frac{1}{n_+} \sum_{x_+ \in X_+} 2\langle \Phi(x), \Phi(x_+) \rangle - \frac{1}{n_+^2} \sum_{x_+ \in X_+} \sum_{x'_+ \in X_+} \langle \Phi(x'_+), \Phi(x_+) \rangle}$$

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

$$d(x, \mu_+) = \sqrt{K(x, x) - \frac{1}{n_+} \sum_{x_+ \in X_+} 2K(x, x_+) - \frac{1}{n_+^2} \sum_{x_+ \in X_+} \sum_{x'_+ \in X_+} K(x'_+, x_+)}$$

Therefore, I will use the previous distance function in the feature space $d(x, \mu_+)$ to express the classifier in terms of the distance to the mean.

The classifier in terms of the sign function and kernel products is then:

$$c(x) = sign(d(x, \mu_-) - d(x, \mu_+))$$

since $d(x, \mu_-)$ is only in terms of kernel products as expressed above, this is the classifier. The above equation tries to capture the intuition "choose the sign of whichever mean of the feature space you are closest to."

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:

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

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

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

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

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

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

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:

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

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