COMS 4771 Fall 2016 Exam 1 solutions

Daniel Hsu

Problem 1

- $(1, A)$: This ERM algorithm only returns homogeneous linear classifiers.
- $(2, C)$: Of the three algorithms, only Online Perceptron might not return a linear separator when one exists.
- $(3, B)$: SVM returns linear separator with largest margin.

Problem 2

(a) **Yes**, the classifiers are the same. For any $w \in \mathbb{R}^d$, $\langle w, x \rangle = \langle \tilde{w}, Ax \rangle$ for all $x \in \mathbb{R}^d$, where $\tilde{w} := (A^\top)^{-1} w$; therefore, linear classifier $f_{w,t}$ is the same as $f_{\tilde{w},t} \circ \phi$.

Ack!! This is incorrect, as I completely overlooked the objective function for SVM. Even though $\tilde{w}$ satisfies the constraints with the transformed data points $(\phi(x_i), y_i)$, it may no longer be the weight vector with the largest margin. This is because now the objective value is $\frac{1}{2}\|\tilde{w}\|_2^2 = \frac{1}{2}\|(A^\top)^{-1} w\|_2^2 = \frac{1}{2} w^\top (A^\top A)^{-1} w$, which is not the same as $\frac{1}{2}\|w\|_2^2$ (unless A is orthogonal).

Because of this slip-up, every student will receive full credit on this problem.

(b) The input is a new point $x_{new} \in \mathbb{R}^d$.

- Compute $\delta_i := K(x_i, x_i) - 2K(x_{new}, x_i) + K(x_{new}, x_{new})$ for each $i \in \{1, 2, \ldots, n\}$. (The $K(x_{new}, x_{new})$ could be omitted.)
- Then compute $i^* := \arg\min_{i \in \{1,2,\ldots,n\}} \delta_i$.
- Finally, return y_{i^*}.

Problem 3

(a) **No**, the error rate of $\hat{f}$ with respect to $\widehat{P}$ cannot be zero. The conditions on the parameters imply that $\widehat{P}(x, y) > 0$ for every $(x, y) \in \{0, 1\}^d \times \{1, 2, \ldots, K\}$. Therefore the Bayes error rate for $\widehat{P}$ is non-zero.

(b) **Yes**, the data set is linearly separable. From (a), we know there is a linear classifier $\hat{f}(x) = \arg\max_{y\in\{1,2,\ldots,K\}}\langle w_y, x\rangle - t_y$, for some $w_y \in \mathbb{R}^d$ and $t_y \in \mathbb{R}$, that satisfies $\hat{f}(x_i) = y_i$ for all $i \in \{1, 2, \ldots, n\}$. Moreover, $\langle w_1, x\rangle - t_1 > \langle w_2, x\rangle - t_2$ if and only if $\langle \tilde{w}, x\rangle > \tilde{t}$ for $\tilde{w} := w_1 - w_2$ and $\tilde{t} := t_1 - t_2$. Therefore, the linear classifier $\tilde{f}$ given by

$$\tilde{f}(x) := \begin{cases} 1 & \text{if } \langle \tilde{w}, x\rangle > \tilde{t}, \\ 2 & \text{otherwise,} \end{cases}$$

is a linear separator for $S_{1,2}$.

(c) The equation is $\sum_{j=1}^{d} \ln \frac{1-\tilde{\mu}_{2,j}}{1-\tilde{\mu}_{1,j}} = 0$ (or $\prod_{j=1}^{d} \frac{1-\tilde{\mu}_{2,j}}{1-\tilde{\mu}_{1,j}} = 1$, or $\ldots$). The Bayes classifier for $\widetilde{P}$ is given by the linear classifier with weight vector $w = (w_1, w_2, \ldots, w_d)$ where $w_j = \ln \frac{\tilde{\mu}_{1,j}(1-\tilde{\mu}_{2,j})}{(1-\tilde{\mu}_{1,j})\tilde{\mu}_{2,j}}$ and threshold $t = \sum_{j=1}^{d} \ln \frac{1-\mu_{2,j}}{1-\mu_{1,j}}$. So the equation given above is the same as $t = 0$.

Problem 4

(a) The Hessian at w is

$$\sum_{i=1}^{n/2} \frac{\exp(y_i\langle w, x_i\rangle)}{(1+\exp(y_i\langle w, x_i\rangle))^2} x_i x_i^\top.$$

The easy way to compute this is to consider the Hessian of each term in the summation, and then to add these matrices together. Considering just the i-th term (and indexing vector entries using $w[\cdot]$ and $x[\cdot]$), we need the Hessian of $f_i(w) := \ln(1 + \exp(-y_i\langle w, x_i\rangle))$. The derivative with respect to $w[j]$ is

$$\frac{\partial f_i(w)}{\partial w[j]} = \frac{\exp(-y_i\langle w, x_i\rangle)}{1+\exp(-y_i\langle w, x_i\rangle)}(-y_i x_i[j]) = \frac{-1}{1+\exp(y_i\langle w, x_i\rangle)} y_i x_i[j].$$

The second derivative with respect to $w[j]$ is

$$\frac{\partial^2 f_i(w)}{\partial w[j]^2} = \frac{\exp(y_i\langle w, x_i\rangle)}{(1+\exp(y_i\langle w, x_i\rangle))^2} x_i[j]^2.$$

The cross derivative with respect to $w[j]$ and $w[k]$ (for $j \neq k$) is

$$\frac{\partial^2 f_i(w)}{\partial w[j]\partial w[k]} = \frac{\exp(y_i\langle w, x_i\rangle)}{(1+\exp(y_i\langle w, x_i\rangle))^2} x_i[j] x_i[k].$$

Therefore the Hessian is

$$\nabla^2 f_i(w) = \frac{\exp(y_i\langle w, x_i\rangle)}{(1+\exp(y_i\langle w, x_i\rangle))^2} x_i x_i^\top.$$

This means that the Hessian for the overall objective function is

$$\sum_{i=1}^{n/2} \frac{\exp(y_i\langle w, x_i\rangle)}{(1+\exp(y_i\langle w, x_i\rangle))^2} x_i x_i^\top.$$

(b) **Yes**, the problem is convex. The objective function is convex because its Hessian at any point $w \in \mathbb{R}^d$ is positive semidefinite (using the hint from part (a)). The i-th constraint (for $i \in \{n/2+1, n/2+2, \ldots, n\}$) can be written as

$$-y_i\langle w, x_i\rangle - \ln(1/2) \leq 0.$$

The left-hand side is an affine function of w, and hence is convex. Therefore the problem is a convex optimization problem.

(c) The new problem is

$$\begin{aligned} \min_{w\in\mathbb{R}^d} \quad & \sum_{i=1}^{n/2} \ln(1+\exp(-y_i\langle w, x_i\rangle)) \\ \text{s.t.} \quad & \|w\|_2^2 - 100 \leq 0. \end{aligned}$$

Yes, the problem is convex. The objective function is convex (as already argued). The left-hand side of the constraint is a convex function of w, because the Hessian matrix is equal to $2I$, which is positive definite. Therefore the new problem is a convex optimization problem.