DS-GA 1003: Machine Learning

March 12, 2019: Midterm Exam (100 Minutes)

Answer the questions in the spaces provided. If you run out of room for an answer, use the blank page at the end of the test. Please **don't miss the last questions**, on the back of the last test page.

Name: ______________________________

NYU NetID: ______________________________

Question	Points	Score
1) Bayes Optimal	7	
2) Risk Decomposition	6	
3) Linear Separability and Loss Functions	6	
4) SVM with Slack Variables	9	
5) Dependent Features	6	
6) RBF Kernel	4	
7) ℓ_2-norm Penalty	6	
Total:	44	

1. (7 points) Consider a binary classification problem. For class $y = 0$, x is sampled from $\{1, 2, 3, 4, 5, 6, 7, 8\}$ with equal probability; for class $y = 1$, x is sampled from $\{7, 8, 9, 10\}$ with equal probability. Assume that both classes are equally likely. Let $f^* : \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\} \to \{0, 1\}$ represent the Bayes prediction function for the given setting under $0 - 1$ loss. Find f^* and calculate the Bayes risk.

Solution: 0-1 loss:

$$l(a, y) = 1(a \neq y) := \begin{cases} 1 & \text{if } a \neq y \\ 0 & \text{otherwise.} \end{cases}$$

Risk:

$$\begin{aligned} R(f) &= \mathbb{E}\left[1(f(x) \neq y)\right] \quad = 0 \cdot \mathbb{P}\left(f(x) = y\right) + 1 \cdot \mathbb{P}\left(f(x) \neq y\right) \\ &= \mathbb{P}\left(f(x) \neq y\right), \end{aligned}$$

which is just the misclassification error rate.

Bayes prediction function is just the assignment to the most likely class,

$$f^*(x) = \underset{c \in \{0,1\}}{argmax}\ p(y = c|x)$$

Therefore:

$$f^*(x) = \begin{cases} 0 & \text{if } x \in \{1, 2, 3, 4, 5, 6\} \\ 1 & \text{if } x \in \{7, 8, 9, 10\} \end{cases}$$

Under $0 - 1$ loss, risk is the probability of mis-classification. $f^*(x)$ mis-classifies points from class 0 occurring in $\{7, 8\}$ as class 1. Hence, bayes risk is

$$\begin{aligned} p(y = 0, x \in \{7, 8\}) &= p(x \in \{7, 8\}|y = 0)p(y = 0) \\ &= \frac{1}{4} \times \frac{1}{2} \\ &= \frac{1}{8} \end{aligned}$$

2. Consider the statistical learning problem for the distribution $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X} = \mathcal{Y} = \mathbf{R}$. A labeled example $(x, y) \in \mathbf{R}^2$ sampled from $\mathcal{D}$ has probability distribution given by $x \sim \mathcal{N}(0, 1)$ and $y|x \sim \mathcal{N}(f^*(x), .1)$, where $f^*(x) = \sum_{i=0}^{5}(i+1)x^i$.

   Let P_k denote the set of all polynomials of degree k on $\mathbf{R}$–that is, the set of all functions of the form $f(x) = \sum_{i=0}^{k} a_i x^i$ for some $a_1, \ldots, a_k \in \mathbf{R}$.

   Let D_m be a training set $(x_1, y_1), \ldots, (x_m, y_m) \in \mathbf{R} \times \mathbf{R}$ drawn i.i.d. from $\mathcal{D}$. We perform empirical risk minimization over a hypothesis space $\mathcal{H}$ for the square loss. That is, we try to find $f \in \mathcal{H}$ minimizing

$$\hat{R}_m(f) = \frac{1}{m}\sum_{i=1}^{m}(f(x) - y)^2$$

   (a) (2 points) If we change the hypothesis space $\mathcal{H}$ from $P_3(x)$ to $P_4(x)$ while keeping the same training set, select **ALL** of the following that **MUST** be true:

   - ☐ Approximation error increases or stays the same.
   - ■ **Approximation error decreases or stays the same.**
   - ☐ Estimation error increases or stays the same.
   - ☐ Bayes risk decreases.

   (b) (2 points) If we change the hypothesis space $\mathcal{H}$ from $P_5(x)$ to $P_6(x)$ while keeping the same training set, select **ALL** of the following that **MUST** be true:

   - ■ **Approximation error stays the same.**
   - ☐ Estimation error stays the same.
   - ☐ Optimization error stays the same.
   - ■ **Bayes risk stays the same.**

   (c) (2 points) If we increase the size of the training set m from 1000 to 5000 while keeping the same hypothesis space $P_5(x)$, select **ALL** of the following that **MUST** be true:

   - ■ **Approximation error stays the same.**
   - ☐ Estimation error decreases or stays the same.
   - ■ **The variance of $\hat{R}_m(f)$ for $f(x) = x^2$ decreases.**
   - ■ **Bayes risk stays the same.**

3. Let D_t denote a training set $(x_1, y_1), \dots, (x_{n_t}, y_{n_t}) \in \mathbf{R}^d \times \{-1, 1\}$ and D_v a validation set $(x_1, y_1), \dots, (x_{n_v}, y_{n_v}) \in \mathbf{R}^d \times \{-1, 1\}$. The training set D_t is linearly separable. Define $J(\theta) = \frac{1}{n_t} \sum_{(x,y) \in D_t} \ell(m)$, where $\ell(m)$ is a margin-based loss function, and m is the margin defined by $m = y(\theta^T x)$.

   We have run an iterative optimization algorithm for 100 steps and attained $\tilde{\theta}$ as our approximate minimizer of $J(\theta)$.

   Denote the training accuracy by $\alpha(D_t) = \frac{1}{n_t} \sum_{(x,y) \in D_t} \mathbf{1}(y\tilde{\theta}^T x > 0)$ and the validation accuracy by $\alpha(D_v) = \frac{1}{n_v} \sum_{(x,y) \in D_v} \mathbf{1}(y\tilde{\theta}^T x > 0)$.

   (a) Answer the following for the logistic loss $\ell(m) = \log(1 + e^{-m})$:

      i. (1 point) __F__ **True or False**: Achieving 100% training accuracy ($\alpha(D_t) = 1$) implies that we have achieved a minimizer of the objective function ($\tilde{\theta} \in \arg\min_\theta J(\theta)$).

      ii. (1 point) __F__ **True or False**: Achieving 100% **validation** accuracy ($\alpha(D_v) = 1$) implies that we have achieved a minimizer of the objective function ($\theta_t \in \arg\min_\theta J(\theta)$).

   (b) Answer the following for the hinge loss $\ell(m) = \max(0, 1 - m)$:

      i. (1 point) __F__ **True or False**: Achieving 100% training accuracy ($\alpha(D_t) = 1$) implies that we have achieved a minimizer of the objective function ($\tilde{\theta} \in \arg\min_\theta J(\theta)$).

      ii. (1 point) __T__ **True or False**: Achieving a minimizer of the objective function ($\tilde{\theta} \in \arg\min_\theta J(\theta)$) implies we have achieved **training** accuracy 100% ($\alpha(D_t) = 1$).

   (c) Answer the following for the perceptron loss $\ell(m) = \max(0, -m)$:

      i. (1 point) __T__ **True or False**: Achieving 100% training accuracy ($\alpha(D_t) = 1$) implies that we have achieved a minimizer of the objective function ($\tilde{\theta} \in \arg\min_\theta J(\theta)$).

      ii. (1 point) __F__ **True or False**: Achieving a minimizer of the objective function ($\tilde{\theta} \in \arg\min_\theta J(\theta)$) implies we have achieved **training** accuracy 100% ($\alpha(D_t) = 1$).

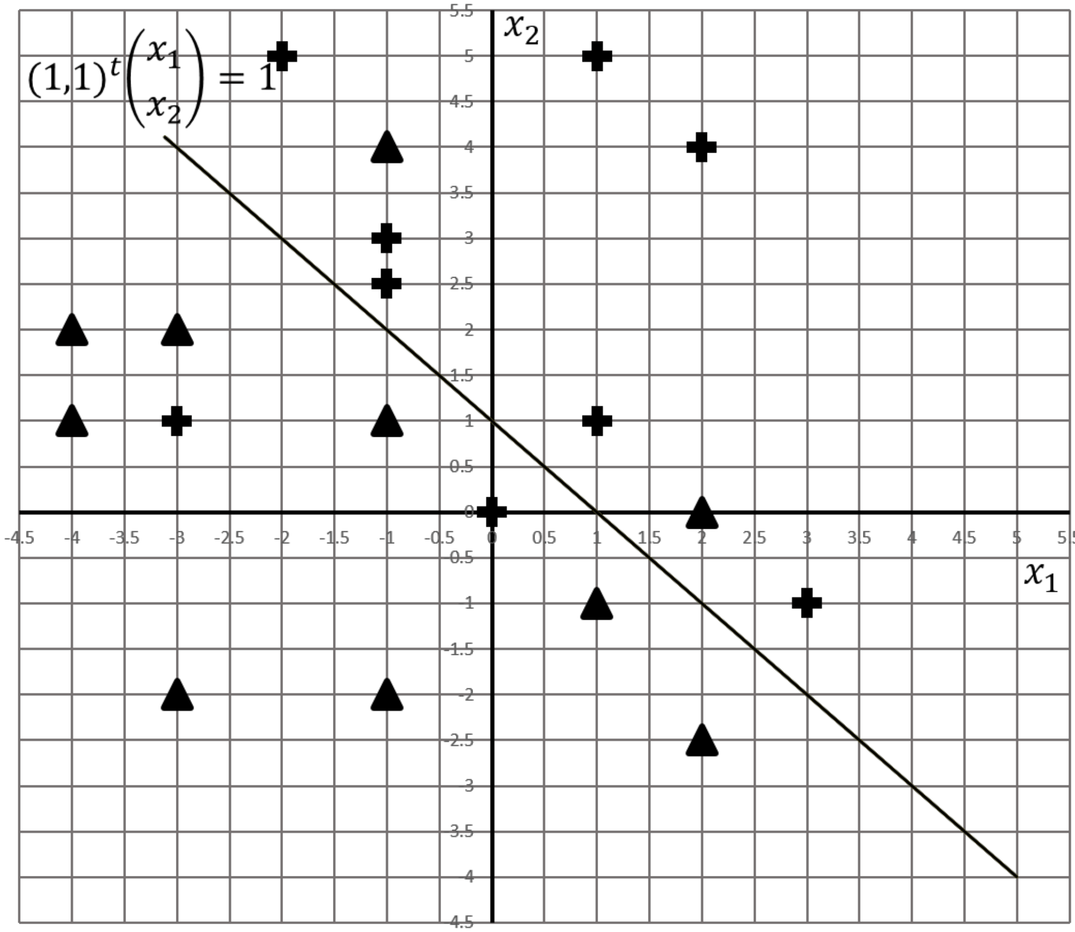

Figure 1: A subset of datapoints from D_m with the decision boundary.

4. Given a dataset $D_m = \{(z_1, y_1), \ldots, (z_m, y_m)\} \in \mathbf{R}^2 \times \{-1, 1\}$ we solve the optimization problem given below to obtain w, b which characterizes the hyperplane which classifies any point $z \in \mathbf{R}^d$ into one of the classes $y = +1$ or $y = -1$ and a number ξ_i for each datapoint $z_i \in D_m$, referred to as slack.

$$\begin{array}{ll} \text{minimize}_{w,b,\xi} & \|w\|_2^2 + \frac{C}{m}\sum_{i=1}^{n} \xi_i \\ \text{subject to} & y_i(w^T z_i - b) \geq 1 - \xi_i \quad \text{for all } i \\ & \xi_i \geq 0 \quad \text{for all } i. \end{array}$$

On solving the optimization problem on D_{100} for some $C \geq 0$, we get that $\hat{w} = (1,1)^T$ and $\hat{b} = 1$. Define $\hat{f}(z) = \hat{w}^T z - \hat{b}$. Figure 1 shows a subset of datapoints from D_m and assume that for all the datapoints $z_i \in D_m$ not shown in Figure 1 we have $y_i \hat{f}(z_i) > 1$. In the figure a label of $+$ represents $y = 1$ and a label of ▲ represents $y = -1$.

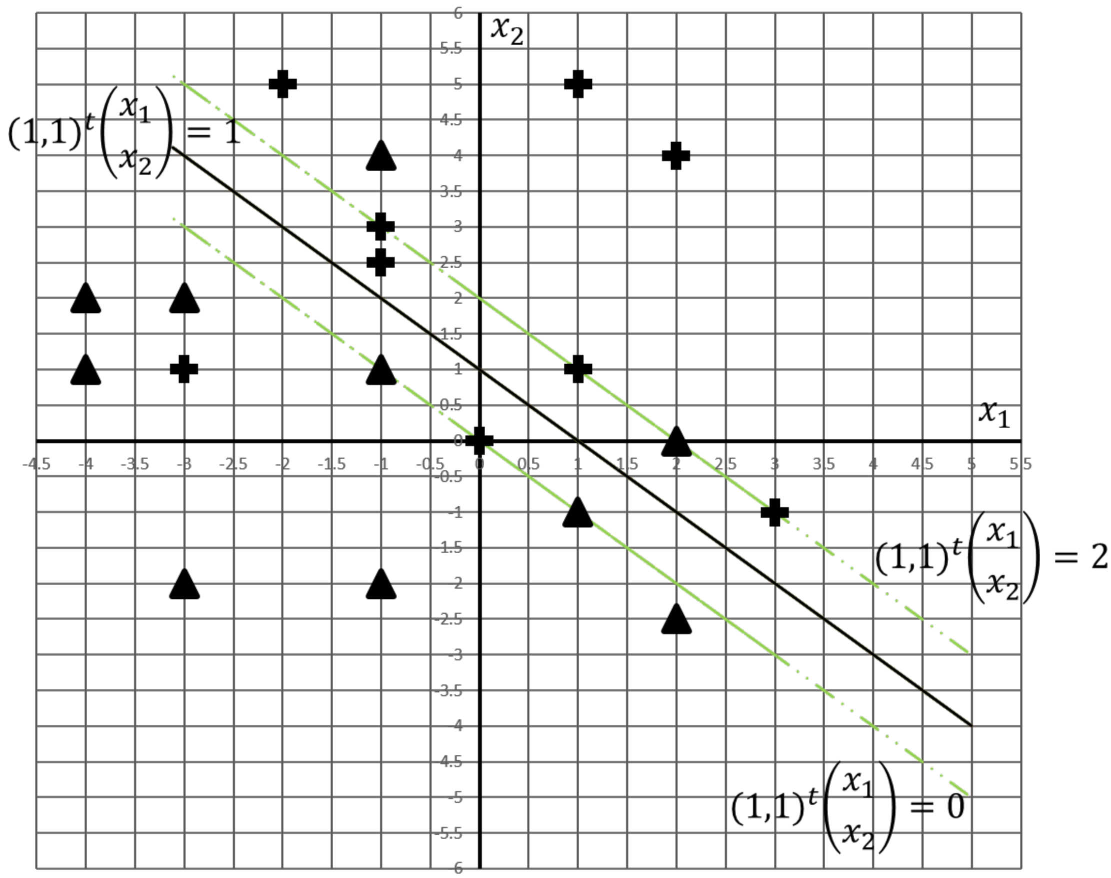

Figure 2: Solution to SVM Question

(a) (2 points) On the graph in figure 1, draw lines to characterize the margin of the classifier $\hat{w}^T z = \hat{b}$. The lines characterizing the margin are defined by $\{z \in \mathbf{R}^2 : \hat{f}(z) = 1\}$ and $\{z \in \mathbf{R}^2 : \hat{f}(z) = -1\}$.

(b) (4 points) Let ξ_{x_1,x_2} denote the slack of the point located at $z = (x_1, x_2)$. For each of the following questions below, fill in the blanks with the best choice from $=, >$ or $<$:

$$\xi_{(2,4)} \underline{\ <\ } \xi_{(2,0)} \qquad \xi_{(-1,1)} \underline{\ =\ } \xi_{(-1,-2)}$$
$$\xi_{(-3,1)} \underline{\ >\ } \xi_{(-1,2.5)} \qquad \xi_{(2,4)} \underline{\ =\ } \xi_{(-1,-2)}$$

(c) From the representer theorem and from duality, we saw that $\hat{w}$ can be expressed as $\hat{w} = \sum_{i=1}^{m} \alpha_i z_i$, where any z_i with $\alpha_i \neq 0$ is called a support vector. The complementary slackness conditions give us the following possibilities for any training example:

1. The example **definitely IS** a support vector.
2. The example **definitely IS NOT** a support vector.
3. We cannot determine from the complementary slackness conditions whether or

not the example is a support vector.

For each of the following training points, select the **ONE** best option from the possibilities above:

i. (1 point) Example at $(2, 4)$ ☐ 1 ■ **2** ☐ 3

ii. (1 point) Example at $(1, 1)$ ☐ 1 ☐ 2 ■ **3**

iii. (1 point) Example at $(2, 0)$ ■ **1** ☐ 2 ☐ 3

5. Let D_n represent a dataset $(x_1, y_1), \dots, (x_n, y_n) \in \mathbf{R}^d \times \mathbf{R}$. The first two dimensions (i.e. features) of every vector x_i are related to each other by scaling: $x_{i1} = sx_{i2}, \forall i = 1, 2, \dots, n$ for some $s \in \mathbf{R}$. Let $X \in \mathbf{R}^{n \times d}$ be the design matrix where the i^{th} row of X contains x_i^T and $\text{rank}(X) = d - 1$ (i.e. there are no other linear dependencies besides the one given). Consider the following objective function for elastic net defined over D_n:

$$J(\theta) = \frac{1}{n} \sum_{i=1}^{n} \left(\theta^T x_i - y_i\right)^2 + \lambda_1 \|\theta\|_1 + \lambda_2 \|\theta\|_2^2$$

(a) Suppose that $|s| \neq 1$. We optimize $J(\theta)$ using subgradient descent. We start the optimization from θ_0 and converge to $\hat{\theta} \in \text{argmin}_{\theta \in \mathbf{R}^d} J(\theta)$. We then restart the optimization from a different point θ_0' and converge to $\hat{\theta}' \in \text{argmin}_{\theta \in \mathbf{R}^d} J(\theta)$. Consider the following possibilities:

1. Must have $\hat{\theta} = \hat{\theta}'$
2. May have $\hat{\theta} \neq \hat{\theta}'$ but must have $J(\hat{\theta}) = J(\hat{\theta}')$
3. May have $\hat{\theta} \neq \hat{\theta}'$ and $J(\hat{\theta}) \neq J(\hat{\theta}')$

For each of the subparts below, select the **ONE** best possibility from the three given above:

i. (1 point) $\lambda_1 = 0, \lambda_2 = 0$ ☐ 1 ■ **2** ☐ 3
ii. (1 point) $\lambda_1 > 0, \lambda_2 = 0$ ■ **1** ☐ 2 ☐ 3
iii. (1 point) $\lambda_1 = 0, \lambda_2 > 0$ ■ **1** ☐ 2 ☐ 3

(b) (3 points) Fix $\lambda_1 = 0$ and $\lambda_2 > 0$. We optimize $J(\theta)$ using stochastic gradient descent, starting from 0, and we attain $\hat{\theta} \in \text{argmin}_{\theta \in \mathbf{R}^d} J(\theta)$. Let $\hat{f}(x) = \hat{\theta}^T x$. Consider a new point $x_t \in \mathbf{R}^d$ such that $x_t^T x_i = 0 \ \forall i = 1, 2, \ldots, n$. Show that $\hat{f}(x_t) = 0$. (This holds for any s, though you should not need to mention s in your answer.)

Solution: First we need to know that $\hat{\theta} = \sum_{i=1}^n \alpha_i x_i$, for some $\alpha \in \mathbf{R}^n$. This follows from the representer theorem or from the fact that in stochastic subgradient descent, we're always taking a step that's a multiple of a data point.

Once we know that θ is in the span of the data, we simply apply it to x_t: $\hat{f}(x_t) = \theta^T x_t = \sum_{i=1}^n \alpha_i x_i^T x_t = 0$.

6. Let $k(x,x') = \exp(-\frac{1}{2\sigma^2}||x - x'||^2_{\mathcal{X}})$, $\sigma > 0$ be the radial basis function (RBF) kernel. By Mercer's theorem, the kernel k corresponds to a feature map $\varphi : \mathcal{X} \to \mathcal{H}$ mapping inputs into an inner product space (actually a Hilbert space). Let $||\cdot||_{\mathcal{H}}$ be the norm in $\mathcal{H}$ and $||\cdot||_{\mathcal{X}}$ be the norm in $\mathcal{X}$.

   (a) (4 points) Show that for any inputs $x_1, x_2, x_3 \in \mathcal{X}$, $||x_2 - x_1||^2_{\mathcal{X}} \leq ||x_3 - x_1||^2_{\mathcal{X}} \implies ||\varphi(x_2) - \varphi(x_1)||^2_{\mathcal{H}} \leq ||\varphi(x_3) - \varphi(x_1)||^2_{\mathcal{H}}$. (Hint: Expand $||\varphi(x) - \varphi(x')||^2$ using inner products, and then derive the conclusion.)

Solution:

$$
\begin{aligned}
\|\varphi(x_2) - \varphi(x_1)\|^2_{\mathcal{H}} &= \langle \varphi(x_2), \varphi(x_2)\rangle + \langle \varphi(x_1), \varphi(x_1)\rangle - 2\langle \varphi(x_2), \varphi(x_1)\rangle \\
&= k(x_2, x_2) + k(x_1, x_1) - 2k(x_2, x_1) \\
&= 1 + 1 - 2\exp(-\frac{1}{2}\|x_1 - x_2\|^2_{\mathcal{X}}) \\
&\leq 1 + 1 - 2\exp(-\frac{1}{2}\|x_1 - x_3\|^2_{\mathcal{X}}) \\
&\leq k(x_3, x_3) + k(x_1, x_1) - 2k(x_3, x_1) \\
&\leq \langle \varphi(x_3), \varphi(x_3)\rangle + \langle \varphi(x_1), \varphi(x_1)\rangle - 2\langle \varphi(x_3), \varphi(x_1)\rangle \\
&\leq \|\varphi(x_3) - \varphi(x_1)\|^2_{\mathcal{H}}
\end{aligned}
$$

7. Consider the regression setting in which $\mathcal{X} = \mathbf{R}^d$, $\mathcal{Y} = \mathbf{R}$, and $\mathcal{A} = \mathbf{R}$ with a linear hypothesis space $\mathcal{F} = \{f(x) = w^T x | w \in \mathbf{R}^d\}$ and the loss function

$$\ell(\hat{y}, y) = (\hat{y} - y)^2$$

where $\hat{y}$ is the action and y is the outcome. Consider the objective function

$$J(w) = \frac{1}{n} \sum_{i=1}^{n} \ell(w^T x_i, y_i) + \lambda \|w\|,$$

where $\|w\| = \sqrt{\sum_{i=1}^{d} w_i^2}$ is the ℓ_2 norm of w.

(a) (4 points) Provide a kernelized objective function $J_k(\alpha) : \mathbf{R}^n \to \mathbf{R}$. You may write your answer in terms of the Gram matrix $K \in \mathbf{R}^{n \times n}$, defined as $K_{ij} = x_i^T x_j$.

Solution: The kernelized objective function is

$$J_k(\alpha) = \frac{1}{n} \sum_{i=1}^{n} \ell((K\alpha)_i, y_i) + \lambda \sqrt{\alpha^T K \alpha}$$

(b) (1 point) __T__ **True or False**: Suppose we use subgradient descent to optimize the objective function and want to find the global minima of the objective function. If we find that there exists a zero subgradient at some step in the subdifferential set, we should stop the subgradient descent immediately.

(c) (1 point) __T__ **True or False**: Let w^* be **any** minimizer of $J(w)$. Then w^* has the form of $w^* = \sum_{i=1}^{n} \alpha_i x_i$.