DSGA-1003 Machine Learning and Computational Statistics

March 1, 2017: Test 1

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 question**, on the back of the last test page.

Name: ____________________

NYU NetID: ____________________

Question	Points	Score
1	4	
2	4	
3	3	
4	1	
5	2	
6	4	
7	4	
8	5	
9	4	
Total:	31	

1. Let $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^d \times \{-1, +1\}$ be a given set of labeled training data.

   (a) (1 point) Give an expression for the (functional, i.e., non-geometric) margin on the data point (x_i, y_i) for an affine score function $f(x) = w^T x + b$ where $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$.

   Solution: $y_i(w^T x_i + b)$

   (b) (1 point) Write the soft-margin SVM objective function in Tikhonov form (i.e., penalty form with no constraints) for the affine hypothesis space

$$H = \{f(x) = w^T x + b \mid w \in \mathbb{R}^d, b \in \mathbb{R}\}.$$

   Solution: The standard formula is

$$J(w) = \|w\|_2^2 + \frac{c}{n}\sum_{i=1}^{n}[1 - y_i(w^T x_i + b)]_+,$$

   where $[z]_+ = z\,\mathbf{1}(z \geq 0)$. Another form is

$$J(w) = \lambda\|w\|_2^2 + \frac{1}{n}\sum_{i=1}^{n}[1 - y_i(w^T x_i + b)]_+.$$

   (c) (2 points) Give an equivalent formulation of the soft-margin SVM optimization problem with a differentiable objective function and affine inequality constraints.

   Solution:

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

2. Consider the variant of the Lasso regression problem given below:

$$\begin{array}{ll} \text{minimize}_{w \in \mathbb{R}^d} & \|Xw - y\|_2^2 \\ \text{subject to} & \|w - v\|_1 \leq r. \end{array}$$

Here $X \in \mathbb{R}^{n \times d}$, $y \in \mathbb{R}^n$, $v \in \mathbb{R}^d$, and $r \in \mathbb{R}_{>0}$ are given.

(a) (1 point) Give the Lagrangian using λ as the dual variable.

> **Solution:**
> $$L(w, \lambda) = \|Xw - y\|_2^2 + \lambda(\|w - v\|_1 - r)$$

(b) (1 point) Prove that strong duality holds. [Recall Slater's condition: For a convex optimization problem, if there exists a $w \in \mathbb{R}^d$ that is strictly feasible, then strong duality holds.]

> **Solution:** Let $w = v$ so that $\|w - v\|_1 = 0 < r$.

(c) (1 point) Complementary slackness conditions specify a relation on the primal and dual optimal variables w^* and λ^*. Write the complementary slackness conditions for this problem.

> **Solution:** $\lambda^*(\|w^* - v\|_1 - r) = 0$

(d) (1 point) Suppose $d = 2$, $r = 1$, $v = (1, 1)^T$, and $\lambda^* = 3$, where λ^* is an optimizing dual variable. Which of the following are possible values of w^*, an optimizing primal variable (select **all** that apply)?

- ☐ $w^* = (0, 0)$
- ☐ $w^* = (1, 1)$
- ■ $w^* = (2, 1)$
- ■ $w^* = (0, 1)$

3. Let $\mathcal{X} = \{1, 2, 3\}$, let $\mathcal{Y} = \{1, 2, 3, 4, 5\}$, and let $\mathcal{A} = \mathcal{Y}$. Suppose the data generating distribution, P, has marginal $X \sim \text{Unif}\{1, 2, 3\}$ and conditional distribution $Y|X = x \sim \text{Unif}\{x, x+1, x+2\}$. Assume we are using the square loss $\ell(a, x) = (a - x)^2$. [Note: Unif denote the uniform distribution on the given set.]

   (a) (1 point) What is the Bayes decision function?

   Solution: $f^*(x) = x + 1$.

   (b) (2 points) What is the Bayes risk?

   Solution:

   $$E[(Y - f^*(X))^2] = E[E[(Y - (X+1))^2|X]] = E[2/3] = \frac{2}{3}.$$

4. (1 point) Which **one** of the following statements is **least plausible** (i.e., probably FALSE) about minibatches for gradient descent.

   - ☐ Improved implementation or improved hardware can allow us to increase the minibatch size and simultaneously reduce convergence time (in seconds).
   - ☐ In general, enlarging the minibatch size (chosen randomly, with replacement) lets us get a better estimate of the full training set gradient.
   - ■ **In general, if we increase the size of our training set by a factor of 1000, then the best minibatch size (with respect to convergence time, in seconds) should also increase by a factor of 1000.**

5. (2 points) Suppose we have a convex objective function (for regularized ERM) and we are currently not at a minimum. Which of the following are **always** descent directions (select **all** that apply)?

   - ☐ Negative of a minibatch gradient.
   - ☐ Negative of a minibatch subgradient.
   - ■ **Negative of the full training set gradient.**
   - ☐ Negative of the full training set subgradient.

6. Let $\mathcal{X} = \mathbb{R}^d$ and let $\mathcal{Y} = \mathcal{A} = \mathbb{R}$. Define the infinite collection of hypothesis spaces $\{\mathcal{F}_r \mid r \geq 0\}$ where

$$\mathcal{F}_r = \{f(x) = w^T x + b \mid w \in \mathbb{R}^d, b \in \mathbb{R}, \|w\|_2 \leq r\}.$$

Define the additional hypothesis space

$$\mathcal{F}_\infty = \{f(x) = w^T x + b \mid w \in \mathbb{R}^d, b \in \mathbb{R}\}.$$

Fix a training set $(x_1, y_1), \ldots, (x_n, y_n)$ where $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$. Throughout, assume we are using some arbitrary fixed loss function ℓ.

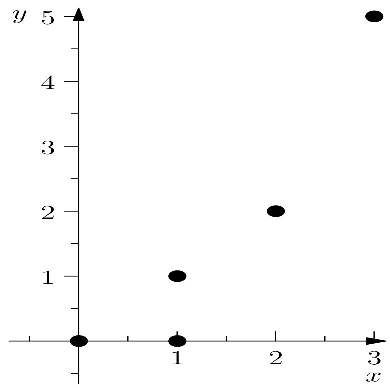
   (a) (1 point) $\underline{\mathcal{F}_\infty}$ Among all hypothesis spaces $\mathcal{F}_r$ for $r \geq 0$, and $\mathcal{F}_\infty$, give a hypothesis space that has empirical risk minimizer with the smallest empirical risk.

   (b) (1 point) $\underline{\mathcal{F}_\infty}$ Among all hypothesis spaces $\mathcal{F}_r$ for $r \geq 0$, and $\mathcal{F}_\infty$, give a hypothesis space that has the lowest approximation error.

   (c) (1 point) __**F**__ **True or False**: Let f_∞ denote the empirical risk minimizer over $\mathcal{F}_\infty$, and let f_c denote the empirical risk minimizer over $\mathcal{F}_c$, where c was chosen by minimizing the loss on a validation set. Then we **always** have $R(f_c) \leq R(f_\infty)$.

   (d) (1 point) __**T**__ **True or False**: Let f_∞ and f_c be as defined previously. Suppose, mistakenly, we reused the training set as the validation set when choosing c. Then we **always** have $\hat{R}(f_c) = \hat{R}(f_\infty)$ (where $\hat{R}$ still refers to the empirical risk on the training set).

7. Let $\mathcal{X} = [0, 1]$ and $\mathcal{Y} = \mathcal{A} = \mathbb{R}$. Suppose you receive the (x, y) data points $(0, 0)$, $(1, 0)$, $(1, 1)$, $(2, 2)$, $(3, 5)$. Throughout assume we are using the $0 - 1$ loss function $\ell(a, y) = \mathbf{1}(a \neq y)$.



(a) (1 point) Suppose we restrict to the hypothesis space $\mathcal{F}_1$ of constant functions. What is the empirical risk minimizer $\hat{f}(x)$?

> **Solution:** $\hat{f}(x) = 0$

(b) (1 point) Suppose we restrict to the hypothesis space $\mathcal{F}_1$ of constant functions. What is $\hat{R}(\hat{f})$, the empirical risk of $\hat{f}$, where $\hat{f}$ is the empirical risk minimizer?

> **Solution:**
> $$\frac{3}{5}$$

(c) (2 points) Suppose we restrict to the hypothesis space $\mathcal{F}_2$ of increasing functions. What is the empirical risk of the associated empirical risk minimizer?

Solution:

$$\frac{1}{5}$$

8. Define the function $h : \mathbb{R} \to \mathbb{R}$ by

$$h(x) = \begin{cases} x^2/2 & \text{if } |x| \le 1, \\ |x| - 1/2 & \text{if } |x| > 1. \end{cases}$$

Consider the objective function

$$J(w) = \frac{1}{n} \sum_{i=1}^{n} h(w^T x_i - y_i) + \lambda \|w\|_2^2$$

where $(x_1, y_1), \dots, (x_n, y_n) \in \mathbb{R}^d \times \mathbb{R}$ and $\lambda > 0$ are given.

(a) (1 point) Which **one** of the following is the **most likely** reason for using h as our loss function instead of the more standard square loss?

☐ The above is the objective for a classification problem, so a different loss is necessary.

■ **The square loss overemphasizes the effect of outliers.**

☐ Using h will enable us to find sparse solutions.

(b) (1 point) We want to minimize $J(w)$ using stochastic gradient descent. Assume the current data point is (x_i, y_i). The SGD step direction is given by $v = -\nabla_w G(w)$, for some function $G(w)$. Give an explicit expression for $G(w)$ in terms of h, λ, and the given data. [Note: You do not have to expand the function h.]

Solution:

$$G(w) = h(w^T x_i - y_i) + \lambda \|w\|_2^2$$

(c) (1 point) Assume $J(w)$ has a minimizer w^*. Give an expression for w^* in terms of a vector $\alpha \in \mathbb{R}^n$ that is guaranteed by the representer theorem. You may use the design matrix $X \in \mathbb{R}^{n \times d}$.

Solution:

$$w^* = \sum_{i=1}^{n} \alpha_i x_i = X^T \alpha$$

(d) (2 points) Let $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a psd kernel, and let $K \in \mathbb{R}^{n \times n}$ denote the matrix with $K_{ij} = k(x_i, x_j)$. Give a kernelized form of the objective J in terms of K. [Hint: $w^T x_i = (Xw)_i$ where $X \in \mathbb{R}^{n \times d}$ is the matrix with ith row x_i^T.]

Solution:

$$J(\alpha) = \frac{1}{n} \sum_{i=1}^{n} h((K\alpha)_i - y_i) + \lambda \alpha^T K \alpha.$$

ONE MORE QUESTION ON THE BACK OF THIS PAGE

9. Consider the following version of the elastic-net objective:

$$J(w) = \frac{1}{n}\|Xw - y\|_2^2 + \lambda_1\|w\|_1 + \lambda_2\|w\|_2^2.$$

Here we have a training set $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^d \times \mathbb{R}$, $X \in \mathbb{R}^{n\times d}$ has x_i^T as its ith row, and $y \in \mathbb{R}^n$ has y_i as its ith coordinate. We fit our data 3 times with the following configurations:

1. Configuration A) $(\lambda_1, \lambda_2) = (0, 0)$
2. Configuration B) $(\lambda_1, \lambda_2) = (5, 0)$
3. Configuration C) $(\lambda_1, \lambda_2) = (0, 5)$

Answer the following questions based on the above information.

(a) For each of the following, state **one** of the configurations that is **most likely** being described. Below w^* represents a minimizer of J.

i. (1 point) __**B**__ w^* has several entries that are 0.

ii. (1 point) __**A**__ The decision function corresponding to w^* has the lowest training error out of all of the configurations.

(b) (2 points) Suppose each data point x has 2 features (x_1, x_2), and that we are using Configuration C. We applied feature normalization which resulted in new scaled features

$$\tilde{x}^T = (\tilde{x}_1, \tilde{x}_2) = (2x_1, x_2/3).$$

This gives the new objective

$$J_s(\tilde{w}) = \frac{1}{n}\|\tilde{X}\tilde{w} - y\|_2^2 + 5\|\tilde{w}\|_2^2$$

which when minimized gives decision function

$$f_{\tilde{w}}(\tilde{x}) = \tilde{w}^T\tilde{x} = 2\tilde{w}_1x_1 + \tilde{w}_2x_2/3.$$

Which **one** of the following unscaled objectives, when minimized, will yield the same decision function? Below we use the unscaled decision function

$$f_w(x) = w_1x_1 + w_2x_2$$

and want $f_w(x) = f_{\tilde{w}}(\tilde{x})$.

☐ $J(w) = \frac{1}{n}\|Xw - y\|_2^2 + 5w_1^2 + 5w_2^2$

■ $J(w) = \frac{1}{n}\|Xw - y\|_2^2 + 5w_1^2/4 + 45w_2^2$

☐ $J(w) = \frac{1}{n}\|Xw - y\|_2^2 + 20w_1^2 + 5w_2^2/9$