DSGA-1003 Machine Learning and Computational Statistics

April 18, 2017: Test 2

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	3	
2	4	
3	2	
4	2	
5	2	
6	4	
7	2	
8	2	
9	2	
10	2	
11	2	
12	3	
13	2	
Total:	32	

1. Recall that the height of a tree is the largest number of edges on any path from the root to any leaf. All answers below should be based on the binary decision trees we covered in class.

   (a) (1 point) In a decision tree of height n, what is the maximum number of leaves possible?

   Solution: 2^n

   (b) (1 point) In a decision tree of height n, what is the minimum number of leaves possible?

   Solution: $n + 1$

   (c) (1 point) Which **ONE** of the following quantities best explains the worst-case running-time when making a prediction using a decision tree?

   - ☐ The number of leaves.
   - ☐ The number of training data points.
   - ☐ The number of features.
   - ■ **The height of the tree.**

2. We are given the dataset $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ where $x_i \in \mathbb{R}^2$ and $y_i \in \{1, 2, 3\}$. Using a one-vs-all methodology we have fit the corresponding score functions $f_{w_i}(x) = w_i^T x$ for $i = 1, 2, 3$ where

$$w_1 = (2, -1), \quad w_2 = (-1, 1), \quad w_3 = (-2, -2).$$

(a) (1 point) To fit each w_i we used a standard soft-margin SVM. If $\mathcal{D} = \big\{\big((0,1),2\big), \big((1,1),1\big), \big((-2,-2),3\big)\big\}$, what dataset was given to the SVM to fit w_3?

Solution:

$$\{\big((0,1),-1\big), \big((1,1),-1\big), \big((-2,-2),1\big)\}$$

(b) (1 point) For each of the following new datapoints x, state which class will be predicted.

i. **2** $x = (-3, 2)$

ii. **1** $x = (1, -1)$

(c) (2 points) We want $\psi : \mathbb{R}^2 \times \{1, 2, 3\} \to \mathbb{R}^D$, for some D, and $\tilde{w} \in \mathbb{R}^D$ so that

$$x \mapsto \arg\max_y \tilde{w}^T \psi(x, y)$$

gives the same classification function as the one-vs-all method described above. Give explicit values for $\tilde{w}$, $\psi(x, 1)$, $\psi(x, 2)$, and $\psi(x, 3)$ for which this is the case. If needed, you may refer to the components of x by $x = (x^1, x^2)$.

Solution:

$$\begin{aligned} \tilde{w} &= (2, -1, -1, 1, -2, -2) \\ \psi(x, 1) &= (x^1, x^2, 0, 0, 0, 0) \\ \psi(x, 2) &= (0, 0, x^1, x^2, 0, 0) \\ \psi(x, 3) &= (0, 0, 0, 0, x^1, x^2) \end{aligned}$$

3. (2 points) Let $\mathcal{D} = \{(1,2),(-2,1),(3,2)\}$ where $\mathcal{X} = \mathcal{Y} = \mathbb{R}$. Give an expression for a prediction function $f : \mathbb{R} \to \mathbb{R}$ that minimizes the square loss over the space of regression stumps (i.e. regression trees of height 1). If you prefer, you may just draw the tree, so long as it contains the same information.

Solution: Many correct answers, but an example is

$$f(x) = \begin{cases} 1 & \text{if } x < 0, \\ 2 & \text{otherwise.} \end{cases}$$

4. (2 points) Select **ALL of the following** that are **TRUE** statements about gradient boosting.

   ☐ A drawback of using gradient boosting is that it cannot be used when the output space is $\mathcal{Y} = \{-1, 1\}$.

   ■ **Empirically, using smaller fixed step sizes leads to better test performance even though training time generally increases.**

   ■ **Randomized feature sampling can be used with gradient boosting to control overfitting.**

   ☐ A drawback of using gradient boosted trees with a custom differentiable loss function ℓ (such as the absolute deviation loss) is that we need a software package that fits regression trees using the custom loss ℓ.

5. (2 points) Consider the dataset $\mathcal{D} = \{(x_1, y_1), (x_2, y_2), (x_3, y_3)\}$ where $\mathcal{X} = \mathcal{Y} = \mathbb{R}$ and

$$(x_1, y_1) = (0, 1), \quad (x_2, y_2) = (3, 2), \quad (x_3, y_3) = (1, -1).$$

Our goal is to use gradient boosting and a software package that fits small regression trees to build our decision function. Suppose we are using the loss function $\ell(y, a) = \log(1 + (y - a)^2)$, that $f_0 \equiv 0$, and we want to compute f_1, the prediction function after one round of gradient boosting. Give the dataset that will be passed into the black box regression tree algorithm to compute f_1.

Solution: Since

$$\partial_a \ell(y, a) = -\frac{2(y - a)}{1 + (y - a)^2}$$

we have

$$-\partial_a \ell(y, 0) = \frac{2y}{1 + y^2}$$

giving

$$\{(0, 2/2), (3, 4/5), (1, -2/2)\} = \{(0, 1), (3, 4/5), (1, -1)\}.$$

6. We are given the dataset $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ where $x_i \in \mathbb{R}^d$ and $y_i \in \{0, 1\}$ and want to fit a conditional probability model such that $y \mid x \sim \text{Bernoulli}(\psi(w^T x))$ where $w \in \mathbb{R}^d$. Recall that the Bernoulli(θ) distribution has probability mass function p given by $p(y) = \theta^y (1-\theta)^{1-y}$ for $y \in \{0, 1\}$ and $p(y) = 0$ otherwise.

   (a) (2 points) Select **ALL of the following** choices for ψ that are allowed given the model.

   ■ **The logistic function** $\psi(t) = 1/(1 + e^{-t})$**.**

   □ The exponential decay function $\psi(t) = e^{-t}$.

   ■ **The standard normal cumulative distribution function**

$$\psi(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-x^2/2}\, dx.$$

   □ The squaring function $\psi(t) = t^2$.

   (b) (2 points) To fit our model to the data, we will minimize the objective function $J(w)$, which is the negative log-likelihood of w given the data $\mathcal{D}$. Give an explicit formula for $J(w)$ as a summation in terms of ψ.

Solution:

$$\begin{aligned} J(w) &= -\log \prod_{i=1}^{n} \psi(w^T x_i)^{y_i} (1 - \psi(w^T x_i))^{1-y_i} \\ &= -\sum_{i=1}^{n} \log \left[\psi(w^T x_i)^{y_i} (1 - \psi(w^T x_i))^{1-y_i} \right] \\ &= -\sum_{i=1}^{n} y_i \log \psi(w^T x_i) + (1 - y_i) \log(1 - \psi(w^T x_i)). \end{aligned}$$

7. (2 points) Select **ALL of the following** that are **TRUE** statements about loss functions.

- ■ **In a conditional probability model, the action space is a set of probability distributions.**
- ■ **Any margin-based loss function $\ell(y, a)$ must be expressible as a function of ya.**
- □ Any loss function $\ell(y, a)$ must take on its minimum value when $y = a$.
- □ When fitting a conditional probability model, the associated loss function $\ell(y, a)$ must be a function of $y - a$.

8. (2 points) Select **ALL of the following** that are **TRUE** statements.

- □ Part of the reason that AdaBoost rarely overfits is that the exponential loss function it uses gives robustness to intrinsic randomness in the labels.
- ■ **AdaBoost follows the FSAM (forward stagewise additive modeling) paradigm. Recall that in each stage of FSAM we solve the optimization problem**

$$(\nu_k, h_k) = \underset{\nu \in \mathbb{R}, h \in \mathcal{H}}{\arg\min} \sum_{i=1}^{n} \ell(y_i, f_{k-1}(x_i) + \nu h(x_i)),$$

  where $\mathcal{H}$ is the base hypothesis space. (This should say Exact AdaBoost, so we are giving everyone 0.5 points for this option, regardless of their answer.)
- □ Random Forests follow the FSAM (forward stagewise additive modeling) paradigm.
- ■ **If AdaBoost can always choose a classifier at each stage with weighted 0-1 loss less than 0.4, then the training loss will converge to 0.**

9. (2 points) Select **ALL of the following** that are **TRUE** statements about decision trees.

- □ With decision trees it is important to center and scale your features as otherwise features with high variance will be selected over low variance features.
- ■ **A classification tree can be used as a conditional probability model by taking the class proportions of the training data in the leaf containing an input x as the estimated probabilities.**
- ■ **Suppose a feature is always positive. Replacing it with its logarithm (in every data point) will have no effect on the tree building process.**
- □ Like square-loss linear regression models, regression trees are heavily influenced by training points with extreme feature values.

10. (2 points) We have a data set $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ where $x_i \in \mathcal{X}$ and $y_i \in \{1, 2, 3\}$. We want to model the conditional distribution of y given x as a categorical distribution with parameters $\theta = (\theta_1, \theta_2, \theta_3) \in [0, 1]^3$, where $\theta_1 + \theta_2 + \theta_3 = 1$. Our model will have the form $\theta = \psi(f_1(x), f_2(x), f_3(x))$, where $f_1, f_2, f_3 : \mathcal{X} \to \mathbb{R}$. Give a valid expression for the function $\psi : \mathbb{R}^3 \to \mathbb{R}^3$. Your answer should have the form $\psi(s_1, s_2, s_3) = (\_\_, \_\_, \_\_)$, where you fill in the 3 blanks.

Solution:

$$\psi(s_1, s_2, s_3) = \left(\frac{e^{s_1}}{\sum_{i=1}^3 e^{s_i}}, \frac{e^{s_2}}{\sum_{i=1}^3 e^{s_i}}, \frac{e^{s_3}}{\sum_{i=1}^3 e^{s_i}} \right) = \frac{1}{\sum_{i=1}^3 e^{s_i}} (e^{s_1}, e^{s_2}, e^{s_3}).$$

11. (2 points) Let $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ with $x_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$. Consider a Gaussian (Bayesian) regression model where $y \mid w, x \sim \mathcal{N}(w^T x, \sigma^2)$ where $\sigma^2 > 0$ is fixed and w has prior $\mathcal{N}(0, \Sigma_0)$ where $\Sigma_0 \succ 0$. Let $X \in \mathbb{R}^{n \times d}$ denote the matrix with x_i^T as its ith row. Using the information above, select **ALL of the following** that are **TRUE** statements.

- ■ **The posterior mode and posterior mean for w are equal.**
- □ By letting $\Sigma_0 = (\sigma^2/\lambda)I$, our posterior mean for w has the same form as is used in Lasso (ℓ_1-regularized) regression.
- ■ **Even though σ is fixed, the posterior predictive variance for y can depend on x.**
- ■ **If X has full rank then the influence of the prior on the posterior for w decreases as σ decreases.**

12. Let $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ with $x_i, y_i \in \mathbb{R}_{>0}$. The exponential distribution with parameter $\lambda > 0$ (denoted ExpDist(λ)) is a distribution on $\mathbb{R}_{>0}$ with PDF given by

$$p(t) = \begin{cases} \lambda e^{-\lambda t} & \text{if } t > 0, \\ 0 & \text{otherwise.} \end{cases}$$

Assume that $y \mid w, x \sim \text{ExpDist}(wx)$ and that w has prior distribution ExpDist(1).

(a) (2 points) Give the PDF $p(w \mid \mathcal{D})$ for the posterior distribution of w given $\mathcal{D}$. You may leave out multiplicative factors that do not depend on w.

Solution:

$$p(w \mid \mathcal{D}) \propto p(w)p(\mathcal{D} \mid w) = e^{-w} \prod_{i=1}^{n} w x_i e^{-w x_i y_i} \propto e^{-w} \prod_{i=1}^{n} w e^{-w x_i y_i} = w^n e^{-w\left(1 + \sum_{i=1}^{n} x_i y_i\right)}.$$

(b) (1 point) Suppose the posterior distribution of w given $\mathcal{D}$ has PDF given by $\varphi_{\mathcal{D}}(w)$. **Give an integral** in terms of $\varphi_{\mathcal{D}}$ for the PDF $p(y \mid x, \mathcal{D})$ of the posterior predictive distribution of y given x and $\mathcal{D}$.

Solution:

$$\begin{aligned} p(y \mid x, \mathcal{D}) &= \int_0^\infty p(y \mid w, x)\varphi_{\mathcal{D}}(w)\, dw \\ &= \int_0^\infty w x e^{-ywx} \varphi_{\mathcal{D}}(w)\, dw. \end{aligned}$$

13. We are given the dataset $\mathcal{D} = \{x_1, \ldots, x_n\}$ drawn from a distribution P, where $x_i \in \mathbb{R}$. Let $\mathcal{D}^{(1)}, \ldots, \mathcal{D}^{(B)}$ be bootstrap samples of size n taken from $\mathcal{D}$.

(a) (1 point) What is the probability that the first data point x_1 does not appear in any of the B bootstrap samples?

Solution:

$$\left[\left(1 - \frac{1}{n}\right)^n\right]^B = \left(1 - \frac{1}{n}\right)^{nB}.$$

(b) (1 point) We would like to estimate the median of the distribution P. Let $\hat{m}_i$ denote the sample median of $\mathcal{D}^{(i)}$. Define

$$\overline{m}_B := \frac{1}{B}\sum_{i=1}^{B} \hat{m}_i.$$

Select **ALL of the following** that are **TRUE** statements.

- ■ **If $\mathcal{D}$ remains fixed and $B \to \infty$ we have $\operatorname{Var} \overline{m}_B \to 0$.**
- □ If $\mathcal{D}$ remains fixed and $B \to \infty$ then the value $\overline{m}_B$ converges to the median of P, the data generating distribution.