DS 100/200: Principles and Techniques of Data Science **Date: April 24, 2020**

Discussion #12

Name:

Logistic Regression

1. Suppose we train a binary classifier on some dataset. Suppose y is the set of true labels, and $\hat{y}$ is the set of predicted labels.

y	0	0	0	0	0	1	1	1	1	1
$\hat{y}$	0	1	1	1	1	1	1	0	0	0

Determine each of the following quantities.

(a) The number of true positives

(b) The number of false negatives

(c) The precision of our classifier. Write your answer as a simplified fraction.

(d) The recall of our classifier. Write your answer as a simplified fraction.

2. You have a classification data set consisting of two (x, y) pairs $(1, 0)$ and $(-1, 1)$.

The covariate vector $\mathbf{x}$ for each pair is a two-element column vector $[1 \quad x]^T$.

You run an algorithm to fit a model for the probability of $Y = 1$ given $\mathbf{x}$:

$$\mathbb{P}\left(Y = 1 \mid \mathbf{x}\right) = \sigma(\mathbf{x}^T \theta)$$

where

$$\sigma(t) = \frac{1}{1 + \exp(-t)}$$

Your algorithm returns $\hat{\theta} = \left[-\frac{1}{2} \quad -\frac{1}{2}\right]^T$

(a) Calculate $\hat{\mathbb{P}}\left(Y = 1 \mid \mathbf{x} = [1 \quad 0]^T\right)$

(b) The empirical risk using log loss (a.k.a., cross-entropy loss) is given by:

$$\begin{aligned} R(\theta) &= \frac{1}{n}\sum_{i=1}^{n} -\log \hat{\mathbb{P}}\left(Y = y_i \mid \mathbf{x_i}\right) \\ &= -\frac{1}{n}\sum_{i=1}^{n} y_i \log \hat{\mathbb{P}}\left(Y = 1 \mid \mathbf{x_i}\right) + (1-y_i)\log \hat{\mathbb{P}}\left(Y = 0 \mid \mathbf{x_i}\right) \end{aligned}$$

And $\hat{\mathbb{P}}\left(Y = 1 \mid \mathbf{x_i}\right) = \sigma(\mathbf{x_i}^T\theta) = \frac{1}{1+\exp(-\mathbf{x_i}^T\theta)} = \frac{\exp(\mathbf{x_i}^T\theta)}{1+\exp(\mathbf{x_i}^T\theta)}$ while $\hat{\mathbb{P}}\left(Y = 0 \mid \mathbf{x_i}\right) = 1 - \hat{\mathbb{P}}\left(Y = 1 \mid \mathbf{x_i}\right) = 1 - \frac{\exp(\mathbf{x_i}^T\theta)}{1+\exp(\mathbf{x_i}^T\theta)} = \frac{1}{1+\exp(\mathbf{x_i}^T\theta)}$. Therefore,

$$\begin{aligned} R(\theta) &= -\frac{1}{n}\sum_{i=1}^{n} y_i \log \frac{\exp(\mathbf{x_i}^T\theta)}{1+\exp(\mathbf{x_i}^T\theta)} + (1-y_i)\log \frac{1}{1+\exp(\mathbf{x_i}^T\theta)} \\ &= -\frac{1}{n}\sum_{i=1}^{n} y_i \mathbf{x}_i^T\theta + \log(\sigma(-\mathbf{x}_i^T\theta)) \end{aligned}$$

Let $\theta = [\theta_0 \quad \theta_1]$. Explicitly write out the empirical risk for the data set $(1, 0)$ and $(-1, 1)$ as a function of θ_0 and θ_1.

(c) Calculate the empirical risk for $\hat{\theta} = \left[-\frac{1}{2} \quad -\frac{1}{2}\right]^T$ and the two observations $(1, 0)$ and $(-1, 1)$.

Decision Trees and Random Forests

3. (a) When creating a decision tree for classification, give two reasons why we might end up having a terminal node that has more than one class.

   (b) Suppose we have a decision tree for classifying the iris data set. Suppose that one terminal decision tree node contains 22 setosas and 13 versicolors. If we're trying to make a prediction and our sequence of yes/no questions leads us to this node, what should we do?

   ◯ A. predict that the class is setosa
   ◯ B. give a probability of setosa = $\sigma(22/35)$
   ◯ C. refuse to make a prediction
   ◯ D. other (describe)

   (c) As mentioned in lecture, we can also use decision trees for regression. Suppose we have the input table given below, where x is our 1 dimensional input value and y is our output value.

x	y
2	4
3	6
4	8
4	10

   i. Draw a valid regression tree for this input.

ii. For your regression tree above, what will your model predict for x = 1?

iii. For your regression tree above, what prediction do you think your model should predict for x = 4?

(d) What techniques can we use to avoid overfitting decision trees?

(e) Suppose we limit the complexity of our decision tree model by setting a maximum possible node depth d, i.e. no new nodes may be created with depth greater than d. What technique should we use to pick d?

(f) What is the advantage of a random forest over a decision tree?

☐ A. lower bias ☐ B. lower variability ☐ C. lower bias and variability ☐ D. none of these