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

# Discussion #12 Solutions

*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

**Solution:** 2

(b) The number of false negatives

**Solution:** 3

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

**Solution:** $\frac{2}{2+4} = \frac{1}{3}$

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

**Solution:** $\frac{2}{2+3} = \frac{2}{5}$

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 $\begin{bmatrix}1 & x\end{bmatrix}^T$.

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 $\hat{\theta} = \begin{bmatrix}-\frac{1}{2} & -\frac{1}{2}\end{bmatrix}^T$

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

**Solution:**

$$\begin{aligned}
\hat{\mathbb{P}}\left(Y = 1 \mid \mathbf{X} = [1 \quad 0]^T\right) &= \sigma\left(\begin{bmatrix}1 & 0\end{bmatrix}\begin{bmatrix}-\frac{1}{2} \\ -\frac{1}{2}\end{bmatrix}\right) \\
&= \sigma\left(1 \times -\frac{1}{2} + 0 \times -\frac{1}{2}\right) \\
&= \sigma\left(-\frac{1}{2}\right) \\
&= \frac{1}{1 + \exp(\frac{1}{2})} \\
&\approx 0.38
\end{aligned}$$

(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 $\theta_0$ and $\theta_1$.

**Solution:**

$$x_i^T\theta = \begin{bmatrix}1 & x_i\end{bmatrix}\begin{bmatrix}\theta_0 \\ \theta_1\end{bmatrix} = \theta_0 + \theta_1 x_i$$

For the data point $(1, 0)$, $\mathbf{x}_i = [1 \quad 1]^T$ and $y_i = 0$, so:

$$y_i \mathbf{x}_i^T\theta = 0$$

$$-\mathbf{x}_i^T\theta = -(\theta_0 + \theta_1 \times 1) = -\theta_0 - \theta_1$$

For the data point $(-1, 1)$:

$$y_i x_i^T\theta = 1 \times (\theta_0 + \theta_1 \times -1) = \theta_0 - \theta_1$$

$$-x_i^T\theta = -(\theta_0 + \theta_1 \times -1) = -\theta_0 + \theta_1$$

We can then write the empirical risk as:

$$\begin{aligned}
R(\theta) &= -\frac{1}{2}\left[(0 + \log\sigma(-\theta_0 - \theta_1)) + (\theta_0 - \theta_1 + \log\sigma(-\theta_0 + \theta_1))\right] \\
&= -\frac{1}{2}\left[\theta_0 - \theta_1 + \log\sigma(-\theta_0 - \theta_1) + \log\sigma(-\theta_0 + \theta_1)\right] \\
&= -\frac{1}{2}\left[\theta_0 - \theta_1 + \log\left(\frac{1}{1+\exp(\theta_0+\theta_1)}\right) + \log\left(\frac{1}{1+\exp(\theta_0-\theta_1)}\right)\right] \\
&= \frac{1}{2}\left[\theta_1 - \theta_0 + \log\left(1+\exp(\theta_0+\theta_1)\right) + \log\left(1+\exp(\theta_0-\theta_1)\right)\right]
\end{aligned}$$

(c) Calculate the empirical risk for $\hat{\theta} = \begin{bmatrix} -\frac{1}{2} & -\frac{1}{2} \end{bmatrix}^T$ and the two observations $(1, 0)$ and $(-1, 1)$.

**Solution:**

$$\begin{aligned}
R(\hat{\theta}) &= \frac{1}{2}\left[\theta_1 - \theta_0 + \log\left(1+\exp(\theta_0+\theta_1)\right) + \log\left(1+\exp(\theta_0-\theta_1)\right)\right] \\
&= \frac{1}{2}\left[-\frac{1}{2} - \left(-\frac{1}{2}\right) + \log\left(1+\exp(-\frac{1}{2}+-\frac{1}{2})\right) + \log\left(1+\exp(-\frac{1}{2}--\frac{1}{2})\right)\right] \\
&= \frac{1}{2}\left[0 + \log\left(1+\exp(-1)\right) + \log\left(1+\exp(0)\right)\right] \\
&= \frac{1}{2}\log(2+2e^{-1})
\end{aligned}$$

# 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.

**Solution:** We could have two input points that are exactly equal but belong to different classes. Or we could have a decision tree that was built using some sort of rule that kept it from becoming too tall, e.g. limiting the maximum depth.

(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?

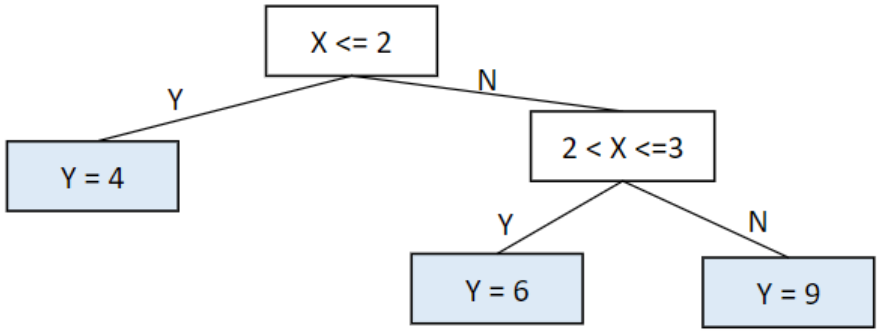
- ◯ A. predict that the class is setosa
- ◯ B. give a probability of setosa = $\sigma(22/35)$
- ◯ C. refuse to make a prediction
- ◯ D. other (describe)

(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.

**Solution:** Here is a potential valid regression tree:




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

**Solution:** 4

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

**Solution:** 9

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

**Solution:** Use a random forest, restrict the maximum tree depth, or prune tree to a particular depth after it is created (do cross-validation to check how much pruning is good).

(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 $d$?

**Solution:** Cross validation. More specifically:

1. Choose a list of arbitrary d values, e.g. $d = (1, 2, 3, ..., N)$
2. Split the data into a training set and a test set
3. Split the training data into K-folds
4. Initialize a vector $r$ to contain the CV risk of each tree model
5. For each depth in $d$, compute the CV-risk of the tree model with depth $d$ and store it in $r$
6. Identify the maximum tree depth value associated with the minimum CV-risk in $r$. This is the optimal model.
7. Train the tree model on the entire training set using the optimal tree depth found in step 5.

(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