DATS 6202	Machine Learning I	Quiz 2
Term 2018-Fall		September 25, 2018

Quiz 2: Solutions

Full Name: ____________________

GWID: ____________________

- DATS 6202, Instructor: Yuxiao Huang

Material Covered

- Logistic regression

Note

- The quiz has 100 points.
- The quiz period is 20 minutes.
- The quiz is closed book and closed notes.
- The quiz is closed electronics (e.g., no laptops, netbooks, OLPCs, tablets, iPads, calculators, cellular phones, iPhones, Nexi, iPods, Zunes, Kindles, Nooks).
- There is only one correct answer for each `Multiple Choice Question`.
- For each `Calculation` question (if there is any), you must show the essential steps. **No mark will be given if only the result is provided**.

1 Multiple Choice Questions (20 points)

1. Which of the following claim is correct about the logit function?

   (a) $\mathrm{logit}(P) = \log\big(\mathrm{odds}(P)\big)$

   (b) $\mathrm{logit}(P) = \mathrm{odds}\big(\log(P)\big)$

   a

2. Which of the following claim is correct about the logistic regression model?

   (a) The parameters can be estimated by maximizing the joint likelihood (the objective function)

   (b) The parameters can be estimated by minimizing the joint likelihood (the objective function)

   a

2 Description and Calculation (80 points)

The logistic regression model can be written as

$$p(y|\mathbf{X}) = \frac{1}{1+e^{-z}} \quad \text{where} \quad z = \sum_{j=0}^{d} \mathrm{w}_j(y) \cdot \mathbf{x}_j. \tag{1}$$

Here,

- y is a class label (e.g., `High`, `Normal`, or `Low`)
- $\mathbf{X}$ is the feature vector
- $p(y|\mathbf{X})$ is the probability of y given $\mathbf{X}$
- d is the number of features
- $\mathbf{x}_j$ is the jth feature (where the dummy feature, $\mathbf{x}_0$, is always 1)
- $\mathrm{w}_j(y)$ is the weight of $\mathbf{x}_j$ with respect to class label y

The rule for updating $\mathrm{w}_j(y)$ (where $0 \leq j \leq d$) can be written as

$$\mathrm{w}_j(y) = \mathrm{w}_j(y) + \Delta\mathrm{w}_j(y) \quad \text{where} \quad \Delta\mathrm{w}_j(y) = \sum_{i=1}^{n} \eta \cdot \Big(f(y_i, y) - P(y|\mathbf{X}_i)\Big) \cdot \mathbf{x}_j. \tag{2}$$

Here,

- $\Delta\mathrm{w}_j(y)$ is the update of $\mathrm{w}_j(y)$
- n (above the $\sum$ symbol) is the number of samples
- η is the learning rate
- y_i is the actual class label of sample i
- y is the predicted class label of sample i (using eq. (1))
- $f(y_i, y)$ is the indicator function, which indicates whether our prediction is correct (i.e., $y_i = y$). That is, $f(y_i, y)$ is 1 when $y_i = y$ and 0 otherwise:

$$f(y_i, y) = \begin{cases} 1, & \text{if } \ y_i = y \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

- $\mathbf{X}_i$ is the feature vector of sample i
- $p(y|\mathbf{X}_i)$ is the probability of y given $\mathbf{X}_i$
- $f(y_i, y) - P(y|\mathbf{X}_i)$ is the error for sample i

1. Explain why η cannot be zero.

   If η were zero, $\Delta \mathrm{w}_j(y)$ would always be zero. As a result, $\mathrm{w}_j(y)$ would not be updated.

2. Explain why η cannot be negative. You should demonstrate that, if η were negative then the updating rule would increase (rather than decrease) the error. You should rely on the following assumption when making your argument.

   - $f(y_i, y) - P(y|\mathbf{X}_i) > 0$ (i.e., the error for sample i is positive)
   - $\mathbf{x}_j > 0$ (i.e., the feature is always positive)

   If η were negative, $\Delta \mathrm{w}_j(y)$ for sample i would be negative (since the assumption says that both the error and the feature is positive). In turn, $\mathrm{w}_j(y)$ would be decreased (based on eq. (2)). Next, first z then $p(y|\mathbf{X})$ would be decreased (based on eq. (1)). Finally the error would be increased.

3. Assume there is one feature $\mathbf{x}_1$ and three class labels (`High`, `Normal`, `Low`) in the data. The parameter values (with respect to each class label) obtained by eq. (2) are as follows:

$$w_0(\texttt{High}) = -1 \quad \text{and} \quad w_1(\texttt{High}) = 1 \tag{4}$$
$$w_0(\texttt{Normal}) = 1 \quad \text{and} \quad w_1(\texttt{Normal}) = -1 \tag{5}$$
$$w_0(\texttt{Low}) = 10 \quad \text{and} \quad w_1(\texttt{Low}) = 10 \tag{6}$$

   Now given a new sample where $\mathrm{x}_1 = 1$:

   - calculate the following probabilities using eq. (1) (where you may assume $e^{-20} \approx 0$):

$$P(\texttt{High}|\mathrm{x}_1 = 1) \tag{7}$$

$$\begin{aligned} P(\texttt{High}|\mathrm{x}_1 = 1) &= \frac{1}{1 + e^{-w_0(\texttt{High}) - w_1(\texttt{High}) \times \mathrm{x}_1}} \\ &= \frac{1}{1 + e^{1 - 1 \times 1}} \\ &= 0.5 \end{aligned}$$

$$P(\texttt{Normal}|\mathrm{x}_1 = 1) \tag{8}$$

$$\begin{aligned} P(\texttt{Normal}|\mathrm{x}_1 = 1) &= \frac{1}{1 + e^{-w_0(\texttt{Normal}) - w_1(\texttt{Normal}) \times \mathrm{x}_1}} \\ &= \frac{1}{1 + e^{-1 + 1 \times 1}} \\ &= 0.5 \end{aligned}$$

$$P(\texttt{Low}|\mathrm{x}_1 = 1) \tag{9}$$

$$\begin{aligned} P(\texttt{Low}|\mathrm{x}_1 = 1) &= \frac{1}{1 + e^{-w_0(\texttt{Low}) - w_1(\texttt{Low}) \times \mathrm{x}_1}} \\ &= \frac{1}{1 + e^{-10-10\times 1}} \\ &\approx 1 \end{aligned}$$

- based on the probabilities above, what is the predicted class label? why?

  The predicted class label is `Low`, since it has the largest probability.

THIS PAGE INTENTIONALLY LEFT BLANK

(You may use it as scratch paper, but *do* submit it as part of your completed exam.)