First Name: Last Name: PIN:

# CIS662 Quiz 5 (11/13/2023)

[Closed-book, open-notes, no communication. Answer each question in the space provided.]

1. **(1 point) How does linear regression differ from logistic regression?**
   Linear regression maps input vectors (data points) to numeric values. Logistic regression maps input vectors to probabilities of class membership (using the logistic function), and a threshold or cutoff value can be used to associate the output with a specific class.

2. (1 point) Consider an ML problem with one input (*x*) and one output (*y*), with training data set $\{(x_1, y_1), (x_2, y_2), \ldots, (x_9, y_9)\}$. One possible algorithm constructs a line $ax+by+c=0$, finding parameters $(a, b, c)$ to minimize the sum of the squared Euclidean distances of points in the training set from the line, i.e., minimizing $\sum_{k=1}^{9} \; (ax_k + by_k + c)^2/(a^2 + b^2)$.

   **How does this differ from linear regression?**

   Linear regression minimizes the sum of the differences between predicted and actual values, $\sum_{k=1}^{9} \; (y_k - (-ax_k - c)/b)^2$, i.e., minimizing $\sum_{k=1}^{9} \; (ax_k + by_k + c)^2 / b^2$ which is different from $\sum_{k=1}^{9} \; (ax_k + by_k + c)^2/(a^2 + b^2)$.

3. (2 points) Consider a linear regression problem in which one of the data points is severely corrupted (e.g., someone enters their annual salary as 1000.00 instead of 100000, or vice versa for their weekly salary). We don't know beforehand which data is corrupted. **How would you address this problem?**

   Any of the following answers would be fine; there may be others which also work:

   - For each attribute of the data, find the anomalous values that are much higher or much lower than the rest. Eliminate them. Try linear regression with the remaining data.
   - Try linear regression with the entire data set. Eliminate points which are farthest from the solution. Try linear regression with the remaining data.
   - Try linear regression with the entire data set. Compute the standard deviation of the distances of points from the result (line or plane) of linear regression. Eliminate points whose distances from the solution are more than 6 times the standard deviation. Try linear regression with the remaining data.

First Name: Last Name: PIN:

# CIS662 Quiz 5 (11/13/2023)

[Closed-book, open-notes, no communication. Answer each question in the space provided.]

1. (2 points) In the dataset for a linear regression problem, a few data points are severely corrupted, but we don't know which data are corrupted. For example, someone may have stated their annual salary to be 999.99 instead of 99999, or vice versa for their weekly salary. **How would you address this problem?**

   Any of the following answers would be fine; there may be others which also work:

   - For each attribute of the data, find the anomalous values that are much higher or much lower than the rest. Eliminate them. Try linear regression with the remaining data.
   - Try linear regression with the entire data set. Eliminate points which are farthest from the solution. Try linear regression with the remaining data.
   - Try linear regression with the entire data set. Compute the standard deviation of the distances of points from the result (line or plane) of linear regression. Eliminate points whose distances from the solution are more than 6 times the standard deviation. Try linear regression with the remaining data.

2. **(1 point) What are the main differences between logistic regression and linear regression?**

   Linear regression maps input vectors (data points) to numeric values. Logistic regression maps input vectors to probabilities of class membership (using the logistic function), and a threshold or cutoff value can be used to associate the output with a specific class.

3. (1 point) Given the training data set $\{(x_1, y_1), (x_2, y_2), \ldots, (x_{99}, y_{99})\}$ where the goal is to predict $y_k$ values from $x_k$ values, one possible algorithm constructs a line $px+qy=r$, finding parameters $(p, q, r)$ to minimize the sum of the squared Euclidean distances of points in the training set from the line, i.e., minimizing $\sum_{k=1}^{99} (px_k + qy_k - r)^2/(p^2 + q^2)$.

   **How does this differ from linear regression?**

   Linear regression minimizes the sum of the differences between predicted and actual values, $\sum_{k=1}^{99} (y_k - (r - px_k)/q)^2$, i.e., minimizing $\sum_{k=1}^{99} (px_k + qy_k - r)^2/q^2$ which is different from $\sum_{k=1}^{99} (px_k + qy_k - r)^2/(p^2 + q^2)$.