EECS545 Lecture 6 Quiz Solutions

1. **Select all that are true.**

   (a) Consider a problem where you want to use a high-dimensional features (where there may be some correlation between the features). Between Naive Bayes and Logistic Regression, Naive Bayes is the better choice.

   (b) Naive Bayes classifier and GDA (Gaussian Discriminant Analysis) are generative models.

   (c) Laplacian smoothing for Naive Bayes avoids zero product for words that show up as only spam / only non-spam

   Solution: (b) and (c).

   (a) is not true: naive Bayes assumes conditional independence of features given class labels. This may be a too strong assumption when there is non-trivial correlation between features.

2. Naive Bayes practice. Consider the following dataset {(spam or not spam, [tokens])} = {(spam, [A, B, B, A]), (not spam, [C, A, B]), (not spam, [B, A, B])}. How many words (vocabulary size M in the lecture) exist in this dataset?

   Solution: $M = 3$ (A, B, C).

3. Continued. Find the naive bayes MLE estimate for P((spam, [C, A, B, B, A])) without laplacian smoothing.

   Solution: $\mu_C^{spam} = 0$, so the entire likelihood is 0.

4. Continued. Find the MLE estimate for P((spam, [C, A, B, B, A])) with laplacian smoothing. We still assume that each token t_i is independent.

Solution:

$$\phi^{spam} = \frac{1}{3} \tag{1}$$

$$\mu_A^{spam} = \frac{2+1}{4+3} = \frac{3}{7} \tag{2}$$

$$\mu_B^{spam} = \frac{2+1}{4+3} = \frac{3}{7} \tag{3}$$

$$\mu_C^{spam} = \frac{0+1}{4+3} = \frac{1}{7} \tag{4}$$

$$P((spam, [\text{C, A, B, B, A}]) = \mu_C^s(\mu_A^s)^2(\mu_B^s)^2\phi^{spam} = \frac{27}{16807} \tag{5}$$