Part 1: Multiple Choice and Short Answer Questions [38 points]

1. **[1 pt] True or False**: One reason that the MAP might be preferred over the MLE is that MLE can have a tendancy to overfit small amounts of data.

   True

2. **[2 pt]** Let X be the result of a coin toss, where $X = 1$ if it comes up heads and $X = 0$ otherwise. The coin has an unknown probability p_1 of coming up heads.
   Suppose that we observe the following sequence of of coin toss outcomes:

   $$(1, 0, 1, 1, 0, 0, 1, 0, 1, 0, 1)$$

   What is the maximum likelihood estimate for p_1?

   0.545

3. **[2 pt]** Now suppose that someone else observes the coin flip (still denoted by X) and tells you Y, the outcome of the flip, but this person only reports the correct result with probability p_2. Suppose we have the following dataset: X-the sequence of actual coin toss outcomes- is:

   $$(1, 0, 1, 1, 0, 0, 1, 0, 1, 0, 1)$$

   Y- the sequence of coin toss outcomes we were told by the other person-is:

   $$(1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1)$$

   What is the maximum likelihood estimate for p_2?

   0.818

4. **[2 pt]** Another person is observing the coin toss too, but the probability for this person to report the correct result depends on the actual outcome of the coin toss. Let $p_{a,b}$ be the probability for that person to report outcome b given that the actual outcome of the coin toss is a , where $a, b \in \{0, 1\}$. Consider the same X and Y values as given in previous question, what is the maximum likelihood estimate for $p_{0,0}$?

   0.8

5. **[4 pt]** Let θ be a random variable with the probability density function:

   $$f(\theta) = \begin{cases} 2\theta, & \text{if } 0 \leq \theta \leq 1 \\ 0, & \text{otherwise.} \end{cases}$$

   Suppose that another random variable Y, conditioning on θ, follows an exponential distribution with $\lambda = 3\theta$. Note that the exponential distribution with parameter λ has a probability density function

   $$f(y) = \begin{cases} \lambda e^{-\lambda y}, & \text{if } y \geq 0, \\ 0, & \text{otherwise.} \end{cases}$$

Find the MAP estimate of θ given $Y = 4$ is observed.

0.166

6. **[1 pt] True or False**: If we choose an incorrect set of parameters for the beta prior of Bernoulli distribution, then the MAP estimate will not converge (as the number of training examples grows toward infinity) to the true value. (here, when we say an 'incorrect' set of parameters for the beta prior, we mean a set of parameters for which the most probable value is different from the true value of the parameter we are trying to estimate.)

   False

7. **[1 pt] True or False**: In case we choose Beta parameters that correspond to a uniform prior, the value of the MAP estimate will be identical to that of the MLE.

   True

8. **[4 pt]** The next two questions refer to the following scenario: suppose that 0.5% people have cancer. Someone decided to take a medical test for cancer. The outcome of the test can either be positive (cancer) or negative (no cancer). The test is not perfect - among people who have cancer, the test comes back positive 96% of the time. Among people who don't have cancer, the test comes back positive 2% of the time. For the following questions, you should assume that the test results are independent of each other, given the true state (*cancer* or *no cancer*).

   What is the probability of a test subject having cancer, given that the subject's test result is positive?

   0.194

9. **[4 pt]** In the same scenario as the previous question, a test subject's first test returned positive, and the subject decided to do a second independent test. The second test returned negative. What is the probability that this subject has cancer?

   0.00975

10. **[1 pt] True or False**: Gaussian Naive Bayes can be used to perfectly classify the training data shown below.

Please refer to the pdf for image of this question.

False

11. **[1 pt] Note**: This question is based on material discussed in Part 2 - Implementing Naïve Bayes of the homework assignment. Please complete Part 2 of this assignment before attempting these questions.

    How many parameters will the model need under the Naïve Bayes assumption, assuming that $P(X_w = x_w|Y = y)$ is a Bernoulli distribution for each w and $P(Y = y)$ is also a Bernoulli distribution? All answers are shown as a function of the vocabulary size V.

    A. V

    B. $2V$

    C. $V + 1$

    D. $2V + 1$

    D. $2V + 1$

12. **[1 pt] Note**: This question is based on material discussed in Part 2 - Implementing Naïve Bayes of the homework assignment. Please complete Part 2 of this assignment before attempting these questions.

    How many parameters (also as a function of V) will the model need if we **do not** make the NB assumption, assuming $P(Y = y)$ is Bernoulli again and all of the features in X have binary labels?

    A. $2V$

    B. 2^V

    C. $2(2^V - 1) + 1$

    D. 2^{2V+1}

    C. $2(2^V - 1) + 1$

13. **[1 pt] Note**: This question is based on material discussed in Part 2 - Implementing Naïve Bayes of the homework assignment. Please complete Part 2 of this assignment before attempting these questions.

    Does the Naïve Bayes assumption hold true for our dataset? Select a valid explanation for your answer.

    A. True. The appearances of each pair of words are not related regardless of review class.

    B. False. The appearances of some common stopwords (say, pronoun *he* and *she*) are dependent in both classes of movie reviews.

    C. True. The number of occurrences for words are not conditionally independent, but the appearances certainly do.

    D. False. For example, *Darth* and *Vader* are unlikely to be independent in both positive and negative reviews.

    D

14. **[1 pt] Note: only one of the answers is correct.** This question is based on material discussed in Part 2 - Implementing Naïve Bayes of the homework assignment. Please complete Part 2 of this assignment before attempting these questions.

    Which of the following statement(s) is/are correct with respect to using stopwords as features?

    A. We can keep stopwords as features. They have no effect on the accuracy of classifier.

    B. Stopwords add value to the dataset which is useful for correctly classifying the document.

    C. Removing stopwords helps in reducing noise/false positives.

    D. All of the above.

    E. None of the above.

    C

15. **[1 pt] Note: only one of the answers is correct.** this is a single choice question. This question is based on material discussed in Part 2 - Implementing Naïve Bayes of the homework assignment. Please complete Part 2 of this assignment before attempting these questions.

    We will experiment with two different parameter settings for our prior over θ_{yw}:

    (a) $\beta_0 = 5$ and $\beta_1 = 7$, and

    (b) $\beta_0 = 7$ and $\beta_1 = 5$.

    Train your classifier with 2 sets of data (`XTrainSmall`,`yTrainSmall`) and (`XTrain`,`yTrain`) with the first parameter setting. Then, use the learned classifiers to classify whether the reviews `XTest` are positive or negative. How do the classification errors compare?

    A. Error is smaller when using `XTrain`,`yTrain`.

    B. Error is smaller when using `XTrainSmall`,`yTrainSmall`.

    C. Errors are equal.

    A. Error is smaller when using XTrain, yTrain

16. **[4 pt] Note**: This question is based on material discussed in Part 2 - Implementing Naïve Bayes of the homework assignment. Please complete Part 2 of this assignment before attempting these questions.

    Train your classifier on the data contained in `XTrain` and `yTrain` with the second parameter setting in the previous problem. Then, use the learned classifier to classify whether the reviews `XTest` are positive or negative. After comparing classification errors produced by classifiers trained by `XTrain` and `yTrain` with 2 parameter settings, which parameter setting was a better choice for the prior on θ_{yw}?

    A. $\beta_0 = 5$ and $\beta_1 = 7$

    B. $\beta_0 = 7$ and $\beta_1 = 5$

    A. $\beta_0 = 5$ and $\beta_1 = 7$

17. **[4 pt] Note**: This question is based on material discussed in Part 2 - Implementing Naïve Bayes of the homework assignment. Please complete Part 2 of this assignment before attempting these questions.

    Consider again the Naïve Bayes classifiers trained with `XTrain` and `yTrain` for both parameter settings. Which of the settings of β_0 and β_1 make more sense if we strongly believe the true value of θ_{yw} lies in the interval $[0.1, 0.3]$?

    A. $\beta_0 = 5$ and $\beta_1 = 7$

    B. $\beta_0 = 7$ and $\beta_1 = 5$

    A. $\beta_0 = 5$ and $\beta_1 = 7$

18. **[0.5 pt] Collaboration Policy Question**: Did you receive any help whatsoever from anyone in solving this assignment? Please answer *yes* or *no*.

    No

19. **[0.5 pt] Collaboration Policy Question**: If you answered *yes* on the previous question, please give full details below (e.g., *Christopher Nolan* explained to me what is asked in Question 3.4).

20. **[0.5 pt] Collaboration Policy Question**: Did you give any help whatsoever to anyone in solving this assignment? Please answer *yes* or *no*.

    No

21. **[0.5 pt] Collaboration Policy Question**: If you answered *yes* on the previous question, please give full details below (e.g., I pointed *Michael Bay* to section 2.3 since he didn't know how to proceed with Question 2).

22. **[0.5 pt] Collaboration Policy Question**: Did you find or come across code that implements any part of this assignment? Please answer *yes* or *no*.

    No

23. **[0.5 pt] Collaboration Policy Question**: If you answered *yes* on the previous question, please give full details below (book & page, URL & location, movies & scene, etc).