Q7. [17 pts] ML: Spam Filter

Pacman has hired you to work on his PacMail email service. You have been given the task of designing a spam detector.

You are given a dataset of emails X, each with labels Y of "spam" or "ham." Here are some examples from the dataset.

Spam Email Ex. 1: "WINNER!! As a valued network customer you have been selected to receive a $900 prize reward!!!"

Spam Email Ex. 2: "We are trying to contact you. Last weekend's draw shows that you won a £1000 prize GUARANTEED!!!"

Ham Email Ex. 1: "Hey! Did you want to grab coffee before the team meeting on Friday?"

Ham Email Ex. 2: "Thank you for attending the talk this morning. I've attached the presentation for you to share with your team. Please let me know if you have any questions."

Your job is to classify the emails in a second dataset, the test dataset, which do not have labels.

(a) [3 pts] Considering only the examples given, which of the following features, in isolation, would be sufficient to classify the examples correctly using a linear classifier?

□ The number of words in the email.
□ The number of times the exclamation point ("!") appears in the email.
□ The number of times "prize" appears in the email.
□ The number of capital letters in the email.
○ None of the above.

(b) [2 pts] Select all true statements about using naive Bayes to solve this problem.

□ We assume that each feature is conditionally independent of the other features, given the label.
□ Given that the prior (class) probabilities are the same, the probability of classifying an email as "spam", given the contents of the email, is proportional to the probability of the contents, given that the email is labeled "spam".
□ Including more features in the model will always increase the test accuracy.
□ Naive Bayes uses the maximum likelihood estimate to compute probabilities in the Bayes net.
○ None of the above.

(c) [3 pts] You want to determine whether naive Bayes or logistic regression is better for your problem. Select all true statements about these two methods.

□ Logistic regression requires fewer learnable parameters than naive Bayes, assuming the same features.
□ Both logistic regression and naive Bayes use the same independence assumption.
□ Both logistic regression and naive Bayes can be used for multi-class classification.
□ Logistic regression models the conditional class distribution $P(Y|W)$ directly, whereas naive Bayes models the joint distribution $P(Y,W)$.
○ None of the above.

You decide to use binary bag-of-words (see definition below) to extract a feature vector from each email in the dataset.

Binary bag-of-words: given a vocabulary of N words, bag-of-words represents a string as an N–element vector, where the value at index i is 1 if word i appears in the string, and 0 otherwise.

The next 3 subparts (d)-(f) are connected. After training a naive Bayes model with **binary bag-of-words** features, you compute the following probability tables for $P(W = w_i | Y = y)$, where w_i is the ith word in your vocabulary. Assume that there are no other words in your model's vocabulary.

	"hey"	"valued"	"team"	"share"
Spam	0.25	0.6	0.3	0.8
Ham	0.4	0.1	0.5	0.3

Now you are tasked with classifying this new email:

"Hey, what time is our team meeting? Can't wait to share with the team!!!"

(d) [1 pt] Fill in the following table with integers corresponding to the bag-of-words vector **w** for the new email. Ignore punctuation and capitalization.

	"hey"	"valued"	"team"	"share"
w				

(e) [3 pts] Compute the probability distribution $P(Y, W = \mathbf{w})$ for this email. You may assume the prior probabilities of each class are equal, i.e. $P(Y = \text{spam}) = P(Y = \text{ham}) = 0.5$. You may ignore words in the sentence that are not present in our model's vocabulary. Write your answer as a single decimal value, rounded to 3 decimal places.

$P(Y = \text{spam}, W = \mathbf{w}) =$ □ $P(Y = \text{ham}, W = \mathbf{w}) =$ □

(f) [2 pts] To get the conditional class distribution from the joint probabilities in part (e), we normalize $P(Y, W = \mathbf{w})$ by dividing it by some Z, such that $P(Y|W = \mathbf{w}) = \frac{1}{Z} P(Y, W = \mathbf{w})$. Write an expression for Z using any of the following terms: $P(Y = \text{ham}, W = \mathbf{w})$, $P(Y = \text{spam}, W = \mathbf{w})$, $P(Y = \text{ham})$, $P(Y = \text{spam})$, and the integer 1.

$Z =$ □

After training the binary bag-of-words model, you find that the test accuracy is still low.

(g) [1 pt] Instead of treating each word as a feature, you decide to use n-grams of words instead. You then split your labeled dataset into a large training set and a small validation set.

Which of the following is the best way of identifying the optimal value of n for your n-gram model?

○ Train the model on the training data using different values of n; select the n with the highest validation accuracy.
○ Train the model on the training data using different values of n; select the n with the highest training accuracy.
○ Select the n that maximizes the number of sequences of n repeated words in the training data.
○ Select n to be the average number of characters per word divided by 2.

(h) [2 pts] You apply Laplace smoothing on the bag-of-words data. Select all true statements.

□ Laplace smoothing for bag-of-words always leads to overfitting.
□ Laplace smoothing is applied by subtracting a constant positive value from each word count.
□ Laplace smoothing eliminates the need for a validation set.
□ Laplace smoothing is only useful for large-vocabulary training datasets.
○ None of the above.

Q7. [13 pts] Machine Learning: Hotdog vs. Not Hotdog

Bob is building a model to classify whether a picture contains a Hotdog or not. He uses two binary features: whether the picture has brown color in it and whether there is red color in it. He collects this training set:

Brown	**Red**	
Color (W_1)	**Color (W_2)**	**Label (y)**
1	0	not hotdog
1	0	not hotdog
1	0	not hotdog
1	1	not hotdog
1	1	hotdog
0	0	hotdog

(a) [5 pts] He first builds a Naive Bayes model. Calculate the following probabilities.

(i) [1 pt] $P(y = \text{hotdog}) = \boxed{\quad}$ $P(y = \text{not hotdog}) = \boxed{\quad}$

(ii) [4 pts] $P(W_1 = 1 \mid y = \text{hotdog}) = \boxed{\quad}$ $P(W_1 = 0 \mid y = \text{hotdog}) = \boxed{\quad}$

$P(W_2 = 1 \mid y = \text{hotdog}) = \boxed{\quad}$ $P(W_2 = 0 \mid y = \text{hotdog}) = \boxed{\quad}$

$P(W_1 = 1 \mid y = \text{not hotdog}) = \boxed{\quad}$ $P(W_1 = 0 \mid y = \text{not hotdog}) = \boxed{\quad}$

$P(W_2 = 1 \mid y = \text{not hotdog}) = \boxed{\quad}$ $P(W_2 = 0 \mid y = \text{not hotdog}) = \boxed{\quad}$

(b) [3 pts] Next, he uses the model to classify three pictures that are from the test set. Fill in the predicted labels in the table.

Test set

Brown	**Red**	**Predicted**
Color (W_1)	**Color (W_2)**	**Label ($\hat{y}$)**
1	1	
0	1	
0	0	

(c) [5 pts] Bob then adds two new examples to his training set, as shown below.

Additional training examples

Brown	**Red**	
Color (W_1)	**Color (W_2)**	**Label (y)**
0	0	not hotdog
0	1	not hotdog

(i) [3 pts] Now re-classify the test set using the new larger training set (hint: don't forget to update the prior for each class with the new dataset).

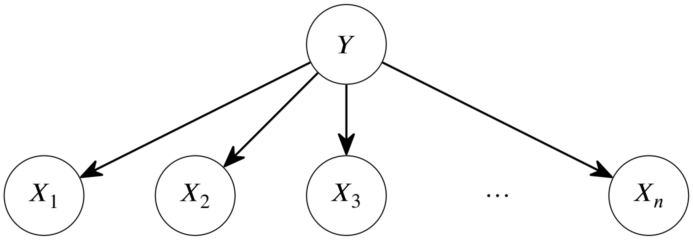
Test set

Brown	**Red**	**Predicted**
Color (W_1)	**Color (W_2)**	**Label ($\hat{y}$)**
1	1	
0	1	
0	0	

(ii) [2 pts] Did any of the predictions change? Why?

Q8. [13 pts] Inverted Naive Bayes

Consider a standard naive Bayes model with $n > 10$ Boolean features $X_1, \dots, X_n$ and a Boolean class variable Y. In this question, Boolean variables have values 0 or 1.



(a) [1 pt] How many parameters do we need to learn in this model (not counting parameters that can be derived by the sum-to-1 rule)?

Example of sum-to-1 rule: Given $P(Y = 0)$, we can compute $P(Y = 1)$ since we know that these two values sum to 1. Therefore, $P(Y = 0)$ and $P(Y = 1)$ only count as one parameter we need to learn.

Your answer should be an expression, possibly in terms of n.

(b) [2 pts] We observe one training example with features $x_1, \dots, x_n$ and class y. Fill in the blanks to derive the likelihood of this training example.

$$L = P(x_1, \dots, x_n, y) = P(Y = 1)^{\textbf{(i)}}\ P(Y = 0)^{\textbf{(ii)}} \prod_{i=1}^{n} P(X_i = 1|y)^{\textbf{(iii)}}\ P(X_i = 0|y)^{\textbf{(iv)}}$$

(i) ○ x ○ $1 - x$ ○ y ○ $1 - y$

(ii) ○ x ○ $1 - x$ ○ y ○ $1 - y$

(iii) ○ x ○ $1 - x$ ○ y ○ $1 - y$

(iv) ○ x ○ $1 - x$ ○ y ○ $1 - y$

(c) [1 pt] Suppose that this training example was missing a feature value, so we only observed $x_2, \dots, x_n$ and class y.

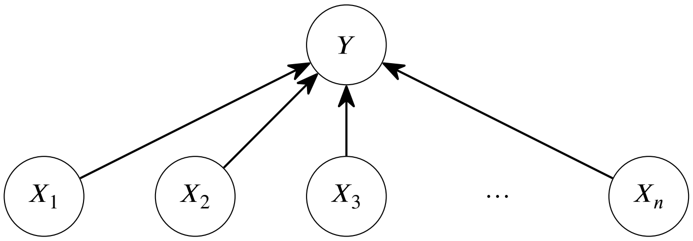
How is the likelihood with a missing feature $L' = P(x_2, \dots, x_n, y)$, related to the original likelihood L?

○ L' is equal to L, but with any terms involving $P(X_1|y)$ dropped.

○ L' is equal to L, with an extra term related to the prior probabilities $P(X_1)$, and any terms involving $P(X_1|y)$ dropped.

○ The correct expression for L' cannot be determined from L.

Suppose we "invert" the naive Bayes model so that the arrows point from the feature variables to the class variable:



(d) [1 pt] How many parameters do we need to learn in this model (not counting parameters that can be derived by the sum-to-1 rule)?

Your answer should be an expression, possibly in terms of n.

For the rest of this question, Boolean variables are represented as positive or negative (e.g. $+y$ and $-y$).

Suppose you have all the parameters from the original naive Bayes model. Using any relevant parameters from the original model, derive the following parameters in the inverted naive Bayes' model, so that the two models represent the same joint distribution.

Your answers should be expressions, possibly in terms of $P(+y)$, $P(-y)$, $P(+x_i|+y)$, $P(-x_i|+y)$, $P(+x_i|-y)$, and $P(-x_i|-y)$, for $1 \leq i \leq n$. Hint: You can also use summations $\sum$ and products $\prod$.

(e) [2 pts] $P(+x_i) =$

(f) [2 pts] $P(+y| + x_1, +x_2, \dots, +x_n) \propto$

(g) [3 pts] Select all true statements.

□ For any set of parameters in the original naive Bayes model, there is some set of parameters in the inverted naive Bayes model that captures the same joint distribution.

□ For any set of parameters in the inverted naive Bayes model, there is some set of parameters in the original naive Bayes model that captures the same joint distribution.

□ The inverted naive Bayes model will typically require far more training data than the original naive Bayes model to achieve the same level of test accuracy.

○ None of the above

(h) [1 pt] What learning behavior should we expect to see with this inverted naive Bayes' model?

○ High training accuracy, high test accuracy

○ High training accuracy, low test accuracy

○ Low training accuracy, high test accuracy

○ Low training accuracy, low test accuracy