Midterm Examination

Tuesday October 28, 9:30am to 10:50am

Your name:

Instructions: Look through the whole exam and answer the questions that you find easiest first. Answer each question in the space below the question, using the backs of the pages for extra space as necessary. If necessary, you may make assumptions that are reasonable, and that do not make a question trivial. If you do make an assumption, state it clearly. This exam is open-book. You may use a calculator.

(Question 1) [20 points]

(a) [4 points] Explain why the following is true: The most probable label of a test example $\langle v_1, v_2, \ldots, v_d \rangle$ is the label $c$ that maximizes

$$P(X_1 = v_1, \ldots, X_d = v_d | C = c)P(C = c).$$

By definition, the label with highest probability is the $c$ that maximizes $P(C = c)P(X_1 = v_1, \ldots, X_d = v_d | C = c)$. Bayes’ rule says that this is equal to

$$\frac{P(X_1 = v_1, \ldots, X_d = v_d | C = c)P(C = c)}{P(X_1 = v_1, \ldots, X_d = v_d)}.$$

The denominator is the same for each alternative value of $c$, so the most probable label is the one that maximizes the numerator.

(b) [5 points] Explain the naive Bayes assumption that lets us simplify the expression $P(X_1 = v_1, \ldots, X_d = v_d | C = c)P(C = c)$.

The naive Bayes assumption is that $P(X_1 = v_1, \ldots, X_d = v_d | C = c) = \prod_{i=1}^{d} P(X_i = v_i | C = c)$.

In words, the assumption is that the random variables $X_1$ to $X_d$ are conditionally independent, conditioned on $C = c$, for every alternative value $c$.

(c) [5 points] Now explain why the following is true: If we make the naive Bayes assumption, then the factor $P(C = c)$ in the expression above is likely to be unimportant.

Typically, because features $X_j$ are correlated (not perfectly independent), the naive Bayes assumption leads to class-conditional probability estimates that are too close to zero:

$$\prod_{i=1}^{d} P(X_i = v_i | C = c) \ll P(X_1 = v_1, \ldots, X_d = v_d | C = c).$$

The predicted label $\hat{c}$ will be whichever choice $c$ gives the largest magnitude for $\prod_{i=1}^{d} P(X_i = v_i | C = c)$. The small values of $\prod_{i=1}^{d} P(X_i = v_i | C = c')$ for $c' \neq \hat{c}$ will outweigh the estimates of $P(C = c')$. 

(d) [6 points] Suppose we have a very large training set (say \( N = 1,000,000 \)) but only two classes \((r = 2)\), only a few features \((d = 4)\) and only a few values per feature \((q = 5)\). Explain how we can train a Bayesian classifier without making the naive Bayes assumption. (Hint: How many parameters do we need to learn, if we don’t make the naive Bayes assumption?)

If we don’t make the naive Bayes assumption, then we need to estimate \( P(X_1 = v_1, \ldots, X_d = v_d | C = c) \) individually for each \( v_1, \ldots, v_d, c \) combination. The number of these combinations is \( q^d r \). Here, \( q^d r = 2^4 \cdot 5^4 = 1250 \). With 1,000,000 training examples we will have 800 examples on average for each of these combinations. This number is large enough to get an accurate estimate

\[
P(X_1 = v_1, \ldots, X_d = v_d | C = c) \approx \frac{\text{count}(X_1 = v_1, \ldots, X_d = v_d, C = c)}{\text{count}(C = c)}.
\]

Of course, for some combinations we will have fewer than 800 training examples, but if these combinations are rare in the training data, then they are also rare in test data, so being less accurate for them is not so important. (We assume that training and test data follow the same distribution.)
(Question 2) [20 points] UCSD has about 26,000 students. The following numbers are approximately true, according to Dr. John Sexton from Counseling and Psychological Services at UCSD. Every year, 1.4% of all students attempt suicide, and there are about 2 actual suicides. Students who have attempted suicide are 543 times more likely to commit suicide successfully in the next year.

(a) [2 points] What percentage of all suicide attempts are successful?
Number of attempts is $0.014 \cdot 26000 = 364$. Success rate is $\frac{2}{364} = 0.55\%$.

(b) [6 points] On average, how many successful suicides per year are by students who did not previously attempt suicide?
Not enough information is given to treat this question from a temporal perspective. So, we will answer it just in terms of percentages. The result will still be useful for deciding where to allocate resources for suicide prevention.
Write $x = p(\text{success} | \neg \text{attempt})$.

\[
p(\text{success} | \text{attempt}) = 543x \\
p(\text{success}) = \frac{2}{26000}
\]

\[
p(\text{success}) = p(\text{success} | \text{attempt})p(\text{attempt}) + p(\text{success} | \neg \text{attempt})p(\neg \text{attempt}) \\
= 543x0.014 + x0.986
\]

\[
\frac{2}{26000} = x(543 \cdot 0.014 + 0.986) \\
x = \frac{\frac{2}{26000}}{543 \cdot 0.014 + 0.986} = 8.957 \cdot 10^{-6}
\]

\[
p(\text{success} \wedge \neg \text{attempt}) = x \cdot p(\neg \text{attempt}) = 8.831 \cdot 10^{-6}
\]

\[
\text{count}(\text{success} \wedge \neg \text{attempt}) = 26000p(\text{success} \wedge \neg \text{attempt}) = 0.230
\]

On average, among the two suicides per year, only 0.23 are by students who did not previously make an attempt. So, it might be sensible to focus prevention efforts on the 364 students who have made previous attempts.

Now, suppose we have a database with detailed information about each UCSD student, with two labels for each person: whether or not s/he actually committed suicide, and whether or not s/he attempted suicide.

(c) [4 points] Explain why it would or would not be useful to train a classifier to distinguish the actual suicides from all other students.

It would not be useful because two is too few examples of the “actual suicide” class. The classifier would overfit the particular training examples of this class. For example, if both happened to be of the same gender, the classifier would predict that the probability of successful suicide was zero (or at least much lower) for the other gender.

(d) [4 points] Suppose you train a classifier to distinguish students who attempt suicide from all other students. Suppose the accuracy of this classifier, measured via cross-validation, is 95%. What can you say about the usefulness of this classifier?

This classifier has lower accuracy than 98.6% as achieved by the classifier that predicts no students ever attempt suicide. However, this classifier may still be useful if it has high precision and/or high recall.

(e) [4 points] Explain why a naive Bayes classifier is likely to be more useful than a 1-nearest neighbor classifier for part (d), even if both classifiers have identical confusion matrices.

With identical confusion matrices, both classifiers have the same accuracy, precision, and recall. However, the naive Bayes classifier can provide more information: its output scores gives a ranking of students from most-likely to least-likely to attempt suicide. Hence, it can be used to focus resources on the students who are at greatest risk. The NN classifier doesn’t provide a useful ranking of test examples.

The naive Bayes classifier will be faster on test examples, but this is not a major benefit, assuming that the NN classifier is fast enough to be usable, which it would be with a training set of cardinality 26,000.

Another benefit of the naive Bayes classifier is that for each feature $X_i$ and feature value $v_i$, the estimates $P(X_i = v_i | C = \text{attempt})$ and $P(X_i = v_i | C = \neg \text{attempt})$ provide information about the predictiveness of this feature value. A nearest-neighbor classifier provides no information about the relative importance of different features or feature values.
1. For use with a nearest neighbor classifier, Euclidean distance and squared Euclidean distance are equivalent.
   True. They lead to the same ranking of neighbors, with the same ties. (One possible non-equivalence is if distances are summed or averaged in order to break ties.)

2. The 3-nearest neighbor classifier is always more accurate than the 2-nearest neighbor classifier.
   False. Two of the three nearest neighbors could be wrong, while tie-breaking makes the 2-nearest neighbor classifier right.

3. With enough training data, the error of a nearest neighbor classifier always goes down to zero.
   False. Some classes could overlap, in which case the Bayes’ error rate is non-zero. Then no classifier can achieve zero error rate.

4. The perceptron algorithm updates the current linear separator if and only if the current training example is misclassified.
   True. This is part of the statement of the algorithm.

5. If the training set is finite and linearly separable, then the perceptron convergence theorem says that the perceptron algorithm will learn a correct linear separator in finite time.
   True. If the training set is finite and linearly separable, then the real values $R$ and $\delta$ needed by the conver-
gence theorem always exist.

6. If the training set is infinite and linearly separable, then the perceptron convergence theorem says that the perceptron algorithm will learn a correct linear separator in finite time.
   *False*. In this case, the norm of training examples may be unbounded, and then $R$ does not exist. And/or, the minimum positive separation $\delta$ may not exist.

7. Cross-validation can reveal overfitting.
   *True*. If accuracy on the training folds is higher than on the test folds, then overfitting is revealed.

8. Overfitting is a danger when learning a classifier, but not when doing unsupervised learning.
   *False*. Informally, overfitting means finding patterns in the data that are spurious because they are true in the training data but false in the test data. It is certainly possible to find spurious patterns in unlabeled data.

9. With $k$-fold cross-validation, larger $k$ is always better.
   *False*. Larger $k$ is slower, and may fail to reveal an issue that would be revealed by smaller $k$. Remember the scenario where every training example is duplicated.

10. Bernoulli and Gaussian distributions are both probability density functions (pdfs).
    *False*. A Gaussian is a pdf but a Bernoulli is a pmf (probability mass function).