CS 4700: Foundations of Artificial Intelligence
Spring 2020
Prep Questions
Quiz 8

1. True / False: Imagine you have a set of data with binary attributes, where within the data for each class every attribute has at least one example that has value 0 for that attribute and another example that has value 1 for that attribute. Naïve Bayes will not generate a 0 probability for any example for any class.

2. Consider using Naïve Bayes on a learning problem with data with 2n attributes x_1, …, x_{2n}.
   a. Imagine that originally the data had only n attributes x_1, …, x_n but due to an error the attributes were duplicated so each x_i and x_{n+i} are identical. However, your learning algorithm doesn't know this.
      i. Would you get different classes assigned to new data if you performed learning on the data with 2n attributes than if you performed learning on the original correct data with n attributes (assume that all ties are broken identically)?
      ii. Would your answer change if Laplace smoothing were used?
   b. Consider a different problem with 2n attributes where Naïve Bayes's conditional independence assumption is false for each pair of attributes x_i and x_{n+i}. In other words, x_1 and x_{n+1} are not conditionally independent, x_2 and x_{n+2} are not conditionally independent, and so on.
      i. Derive a new form of the Naïve Bayes learning rule so that you estimate $P(c_k|x_{test})$ using the product of $P(x_{test,i}, x_{test,n+i} |c_k)$ terms instead of the usual $P(x_{test,i} |c_k)$ terms. Recall Bayes Rule:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

      ii. How would you estimate each $P(x_{test,i}, x_{test,n+i} |c_k)$ term from the training data? Assume you are using Laplace smoothing to compute this estimate.

3. Consider the following set of training data. Each row is an example. Each example is described by three features, x_1, x_2, and x_3, and falls into one of two categories, + or –.

x_1	x_2	x_3	Label
1	1	0	–
0	0	0	–
0	1	0	–
0	0	1	+
1	1	1	+
0	1	1	–
1	0	1	–

   a. What category would Naïve Bayes assign to an example with x_1= 1, x_2= 0, and x_3= 0 if Laplace smoothing is not used? Show your work.
   b. What category would be assigned to the example if Laplace smoothing is used for the features with α=1? Show your work.

1. True / False: Imagine you have a set of data with binary attributes, where within the data for each class every attribute has at least one example that has value 0 for that attribute and another example that has value 1 for that attribute. Naïve Bayes will not generate a 0 probability for any example for any class.

   True. We're computing $\text{argmax}_{c\in C} P(c)\prod_{i=1}^{n} P(x_i|c)$. The conditions in the statement above imply that each of the probabilities that are being multiplied together are nonzero, which means their product will similarly be nonzero.

2. Consider using Naïve Bayes on a learning problem with data with 2n attributes $x_1, \ldots, x_{2n}$.
   a. Imagine that originally the data had only n attributes $x_1, \ldots, x_n$ but due to an error the attributes were duplicated so each x_i and x_{n+i} are identical. However, your learning algorithm doesn't know this.
      i. Would you get different classes assigned to new data if you performed learning on the data with 2n attributes than if you performed learning on the original correct data with n attributes (assume that all ties are broken identically)?

      Yes. In the case of the original n attributes we'd be doing

$$\underset{c\in C}{\text{argmax}}\, P(c)\prod_{i=1}^{n} P(x_i|c)$$

      For 2n attributes it's

$$\underset{c\in C}{\text{argmax}}\, P(c)\cdot\prod_{i=1}^{2n} P(x_i|c) = \underset{c\in C}{\text{argmax}}\, P(c)\cdot\prod_{i=1}^{n} P(x_i|c)\cdot\prod_{i=n+1}^{2n} P(x_i|c)$$

$$= \underset{c\in C}{\text{argmax}}\, P(c)\cdot\left(\prod_{i=1}^{n} P(x_i|c)\right)^2$$

      If they give different labels (let's say for the first it's c_1 and for the second c_2) then we know that

$$P(c_1)\prod_{i=1}^{n} P(x_i|c_1) > P(c_2)\prod_{i=1}^{n} P(x_i|c_2)$$

      for the n-attribute case, or in other words

$$\frac{P(c_1)}{P(c_2)} > \frac{\prod_{i=1}^{n} P(x_i|c_2)}{\prod_{i=1}^{n} P(x_i|c_1)}$$

      Similarly, for the 2n-attribute case since c_2 is the assigned class we know that

$$P(c_2)\left(\prod_{i=1}^{n} P(x_i|c_2)\right)^2 > P(c_1)\left(\prod_{i=1}^{n} P(x_i|c_1)\right)^2$$

      or in other words

$$\frac{P(c_1)}{P(c_2)} < \frac{\left(\prod_{i=1}^{n} P(x_i|c_2)\right)^2}{\left(\prod_{i=1}^{n} P(x_i|c_1)\right)^2}$$

      Combined this means

$$\frac{\prod_{i=1}^{n} P(x_i|c_2)}{\prod_{i=1}^{n} P(x_i|c_1)} < \left(\frac{\prod_{i=1}^{n} P(x_i|c_2)}{\prod_{i=1}^{n} P(x_i|c_1)}\right)^2$$

or in other words

$$\prod_{i=1}^{n} P(x_i|c_1) < \prod_{i=1}^{n} P(x_i|c_2)$$

It is thus possible to have different labels, where the final above inequality would be true.

ii. Would your answer change if Laplace smoothing were used?

No. The answer above didn't depend on how the probabilities are estimated, just that

$$\prod_{i=1}^{n} P(x_i|c_1) < \prod_{i=1}^{n} P(x_i|c_2)$$

b. Consider a different problem with 2n attributes where Naïve Bayes's conditional independence assumption is false for each pair of attributes x_i and x_{n+i}. In other words, x_1 and x_{n+1} are not conditionally independent, x_2 and x_{n+2} are not conditionally independent, and so on.

i. Derive a new form of the Naïve Bayes learning rule so that you estimate $P(c_k|x_{test})$ using the product of $P(x_{test,i}, x_{test,n+i}\ |c_k)$ terms instead of the usual $P(x_{test,i}\ |c_k)$ terms. Recall Bayes Rule:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

We would treat each *pair* of conditionally dependent variables as if they were a single variable, conditionally independent from all the other such pairs:

$$\underset{c \in C}{\operatorname{argmax}} P(c) \prod_{i=1}^{n} P\left(x_{test,i}, x_{test,n+i} \middle| c\right)$$

ii. How would you estimate each $P(x_{test,i}, x_{test,n+i}\ |c_k)$ term from the training data? Assume you are using Laplace smoothing to compute this estimate.

Recall that the set of values that an attribute x_i can take on is V_i, and the number of values would be $|V_i|$. Laplace smoothing on a single variable makes it as if the values it can take on are uniformly distributed. Here we would want to treat each *pair* of values that the two variables might take on as uniformly distributed – every combination of a value from V_i and V_{n+i} should be equally probable, or in other words $\frac{1}{|V_i||V_{n+i}|}$. This uses $|V_i||V_{n+i}|$ in the same way we had previously used $|V_i|$ in Laplace smoothing, meaning that we would estimate $P(x_{test,i}, x_{test,n+i}\ |c)$ by

$$\frac{\text{\# of examples with class c that have the values of } x_{test,i}, x_{test,n+I} + k}{\text{\# of examples with class } c_k + k \cdot |V_i||V_{n+i}|}$$

3. Consider the following set of training data. Each row is an example. Each example is described by three features, x_1, x_2, and x_3, and falls into one of two categories, + or –.

x_1	x_2	x_3	Label
1	1	0	–
0	0	0	–
0	1	0	–

0	0	1	+
1	1	1	+
0	1	1	–
1	0	1	–

a. What category would Naïve Bayes assign to an example with x_1= 1, x_2= 0, and x_3= 0 if Laplace smoothing is not used? Show your work.

+: P(c) = 2/7, P(x1=1|c)=1/2, P(x2=0|c)=1/2, P(x3=0|c)=0
Multiply them together and you get 0.

–: P(c) = 5/7, P(x1=1|c)=2/5, P(x2=0|c)=2/5, P(x3=0|c)=3/5
Multiply them together and you get 12/175.

Thus predict -.

b. What category would be assigned to the example if Laplace smoothing is used for the features with α=1? Show your work.

+: 2/7 * 2/4 * 2/4 * 1/4 = 1/56
–: (1) 5/7 * 3/7 * 3/7 * 4/7

Thus predict -.