CSE 312: Foundations of Computing II
Quiz Section #4: Conditional Probability, Random Variables, Naive Bayes
(solutions)

Review: Main Theorems and Concepts

Conditional Probability: $\mathbb{P}(A|B) = \dfrac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$

Independence: Events E and F are independent iff $\mathbb{P}(E \cap F) = \mathbb{P}(E)\mathbb{P}(F)$, or equivalently $\mathbb{P}(F) = \mathbb{P}(F|E)$, or equivalently $\mathbb{P}(E) = \mathbb{P}(E|F)$

Conditional Independence: Let E, F, G be events, and $\mathbb{P}(G) > 0$. E and F are conditionally independent given G iff $\mathbb{P}(E \cap F \mid G) = \mathbb{P}(E \mid G)\mathbb{P}(F \mid G)$, or equivalently $\mathbb{P}(F|E, G) = \mathbb{P}(F|G)$, or equivalently $\mathbb{P}(E|F, G) = \mathbb{P}(E|G)$

Bayes Theorem: $\mathbb{P}(A|B) = \dfrac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)}$

Partition: Nonempty events $E_1, \ldots, E_n$ partition the sample space Ω iff

- $E_1, \ldots, E_n$ are exhaustive: $E_1 \cup E_2 \cup \cdots \cup E_n = \bigcup_{i=1}^n E_i = \Omega$, and
- $E_1, \ldots, E_n$ are pairwise mutually exclusive: $\forall i \neq j, E_i \cap E_j = \emptyset$
  - Note that for any event A (with $A \neq \emptyset, A \neq \Omega$): A and A^C partition Ω

Law of Total Probability (LTP): Suppose $A_1, \ldots, A_n$ partition Ω and let B be any event. Then

$\mathbb{P}(B) = \sum_{i=1}^n \mathbb{P}(B \cap A_i) = \sum_{i=1}^n \mathbb{P}(B \mid A_i)\mathbb{P}(A_i)$

Bayes Theorem with LTP: Suppose $A_1, \ldots, A_n$ partition Ω and let A, B be any events with $\mathbb{P}(A), \mathbb{P}(B) > 0$.

Then $\mathbb{P}(A_1|B) = \dfrac{\mathbb{P}(B \mid A_1)\mathbb{P}(A_1)}{\sum_{i=1}^n \mathbb{P}(B \mid A_i)\mathbb{P}(A_i)}$. In particular,

$\mathbb{P}(A|B) = \dfrac{\mathbb{P}(B \mid A)\mathbb{P}(A)}{\mathbb{P}(B \mid A)\mathbb{P}(A) + \mathbb{P}(B \mid A^C)\mathbb{P}(A^C)}$

Chain Rule: Suppose $A_1, \ldots, A_n$ are events. Then

$$\mathbb{P}(A_1 \cap \ldots \cap A_n) = \mathbb{P}(A_1)\mathbb{P}(A_2 \mid A_1)\mathbb{P}(A_3 \mid A_1 \cap A_2) \ldots \mathbb{P}(A_n \mid A_1 \cap \ldots \cap A_{n-1})$$

Random Variable (rv): A numeric function $X : \Omega \to \mathbb{R}$ of the outcome.

Range/Support: The support/range of a random variable X, denoted Ω_X, is the set of all possible values that X can take on.

Discrete Random Variable (drv): A random variable taking on a countable (either finite or countably infinite) number of possible values.

Probability Mass Function (pmf) for a discrete random variable X: a function $p_X : \Omega_X \to [0,1]$ with $p_X(x) = \mathbb{P}(X = x)$ that maps possible values of a discrete random variable to the probability of that value happening, such that $\sum_x p_X(x) = 1$.

Exercises

1. Suppose there are three possible teachers for CSE 312: Martin Tompa, Anna Karlin, and Anup Rao. Suppose the ratio of grades $A : B : C : D : F$ for Martin's class is $1 : 2 : 3 : 4 : 5$, for Anna's class is $3 : 4 : 5 : 1 : 2$, and for Anup's class is $5 : 4 : 3 : 2 : 1$. Suppose you are assigned a grade randomly according to the given ratios when you take a class from one of these professors, irrespective of your performance. Furthermore, suppose Martin teaches your class with probability $\frac{1}{2}$ and Anna and Anup have an equal chance of teaching if Martin isn't. What is the probability you had Martin, given that you received an A? Compare this to the unconditional probability that you had Martin.

   Let T, K, R be the events you take 312 from Tompa, Karlin, and Ruzzo, respectively. Let the letter grades be events themselves.

$$\mathbb{P}(T|A) = \frac{\mathbb{P}(A|T)\mathbb{P}(T)}{\mathbb{P}(A|T)\mathbb{P}(T) + \mathbb{P}(A|K)\mathbb{P}(K) + \mathbb{P}(A|R)\mathbb{P}(R)} = \frac{\frac{1}{15}\cdot\frac{1}{2}}{\frac{1}{15}\cdot\frac{1}{2} + \frac{3}{15}\cdot\frac{1}{4} + \frac{5}{15}\cdot\frac{1}{4}}$$

$$= \frac{\frac{1}{2}}{\frac{1}{2} + \frac{3}{4} + \frac{5}{4}} = \frac{2}{2+3+5} = \frac{2}{10} = \boxed{\frac{1}{5}}$$

2. Suppose we have a coin with probability p of heads. Suppose we flip this coin n times independently. Let X be the number of heads that we observe. What is $\mathbb{P}(X = k)$, for $k = 0, \ldots n$? Verify that $\sum_{k=0}^{n} \mathbb{P}(X = k) = 1$, as it should.

$$\mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

   For a given sequence with exactly k heads, the probability of that sequence is $p^k(1-p)^{n-k}$. However, there are $\binom{n}{k}$ such sequences, so the probability of exactly k heads is $\binom{n}{k}p^k(1-p)^{n-k}$.

$$\sum_{k=0}^{n} \mathbb{P}(X = k) = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} = (p + (1-p))^n = 1$$

   The middle equality uses the Binomial Theorem.

3. Suppose we have a coin with probability p of heads. Suppose we flip this coin until we flip a head for the first time. Let X be the number of times we flip the coin *up to and including* the first head. What is $\mathbb{P}(X = k)$, for $k \in \mathbb{N}$? Verify that $\sum_{k=1}^{\infty} \mathbb{P}(X = k) = 1$, as it should.

$$\mathbb{P}(X = k) = (1-p)^{k-1} p$$

   If the k^{th} flip is our first head, the first $k-1$ must be tails. Then the k^{th} flip must be a head.

$$\sum_{k=1}^{\infty} \mathbb{P}(X = k) = \sum_{k=1}^{\infty} (1-p)^{k-1} p = p \sum_{j=0}^{\infty} (1-p)^j = \frac{p}{1-(1-p)} = 1$$

4. A frog starts on a 1-dimensional number line at 0. At each second, independently, the frog takes a unit step right with probability p_1, to the left with probability p_2, and doesn't move with probability p_3, where $p_1 + p_2 + p_3 = 1$. After 2 seconds, let X be the location of the frog. Find $p_X(k)$, the probability mass function for X. Find $p_Y(k)$, the probability mass function for $Y = |X|$.

   Let L be a left step, R be a right step, and N be no step.

$$\begin{aligned}
\mathbb{P}(X = -2) &= \mathbb{P}(\text{LL}) = p_2^2 \\
\mathbb{P}(X = 2) &= \mathbb{P}(\text{RR}) = p_1^2 \\
\mathbb{P}(X = 1) &= \mathbb{P}(RN \cup NR) = 2p_1p_3 \\
\mathbb{P}(X = -1) &= \mathbb{P}(LN \cup NL) = 2p_2p_3 \\
\mathbb{P}(X = 0) &= \mathbb{P}(NN \cup LR \cup RL) = p_3^2 + 2p_1p_2 \\
p_X(k) &= \begin{cases} p_2^2, & k = -2 \\ 2p_2p_3, & k = -1 \\ p_3^2 + 2p_1p_2, & k = 0 \\ 2p_1p_3, & k = 1 \\ p_1^2, & k = 2 \end{cases} \\
p_Y(k) &= \begin{cases} p_3^2 + 2p_1p_2, & k = 0 \\ 2p_3(p_1 + p_2), & k = 1 \\ p_1^2 + p_2^2, & k = 2 \end{cases}
\end{aligned}$$

5. Corrupted by their power, the judges running the popular game show America's Next Top Mathematician have been taking bribes from many of the contestants. During each of two episodes, a given contestant is either allowed to stay on the show or is kicked off. If the contestant has been bribing the judges, she will be allowed to stay with probability 1. If the contestant has not been bribing the judges, she will be allowed to stay with probability 1/3, independent of what happens in earlier episodes. Suppose that 1/4 of the contestants have been bribing the judges. The same contestants bribe the judges in both rounds.

   (a) If you pick a random contestant, what is the probability that she is allowed to stay during the first episode?

   Let S_i be the event that she stayed during the i-th episode. By the Law of Total Probability,

$$\mathbb{P}(S_1) = \frac{1}{4} \cdot 1 + \frac{3}{4} \cdot \frac{1}{3} = \frac{1}{2}$$

(b) If you pick a random contestant, what is the probability that she is allowed to stay during both episodes?

By the Law of Total Probability,

$$\mathbb{P}(S_1 \cap S_2) = \frac{1}{4} \cdot 1 \cdot 1 + \frac{3}{4} \cdot \frac{1}{3} \cdot \frac{1}{3} = \frac{1}{3}$$

(c) If you pick a random contestant who was allowed to stay during the first episode, what is the probability that she gets kicked off during the second episode?

By the definition of conditional probability and the Law of Total Probability,

$$\mathbb{P}(\overline{S_2} \mid S_1) = \frac{\mathbb{P}(S_1 \cap \overline{S_2})}{\mathbb{P}(S_1)} = \frac{\frac{1}{4} \cdot 1 \cdot 0 + \frac{3}{4} \cdot \frac{1}{3} \cdot \frac{2}{3}}{\frac{1}{2}} = \frac{1/6}{1/2} = \frac{1}{3}$$

(d) If you pick a random contestant who was allowed to stay during the first episode, what is the probability that she was bribing the judges?

Let B be the event that she bribed the judges. By Bayes' Theorem,

$$\mathbb{P}(B \mid S_1) = \frac{\mathbb{P}(S_1 \mid B)\mathbb{P}(B)}{\mathbb{P}(S_1)} = \frac{1 \cdot \frac{1}{4}}{\frac{1}{2}} = \frac{1}{2}$$

6. Questions about the Naive Bayes Classifier:

(a) Naive Bayes assumes conditional independence of words in an email, given that we know the label (ham or spam) of the email. Why is that assumption necessary to make Naive Bayes work?

Without assuming conditional independence, the chain rule gives rise to factors such as $P(x_1 \mid x_2, \ldots, x_n, S)$ that have so many conditions that it's impossible to estimate its probability. Think about what that term is trying to calculate: what is the probability that the word x_1 occurs in a spam email, given that $x_2, \ldots x_n$ also occur? Unless there are several spam emails in the training data with all the words $x_2, \ldots x_n$ in them, the training procedure can provide no useful estimate of this probability.

(b) Is the conditional independence assumption actually true in the real world? That is, are the occurrences of words in an email independent of each other, if we know the label of the email? Explain.

No. Certain words tend to occur together, such as "top" and "secret", and "viagra" and "man".

(c) Do you expect the Naive Bayes Classifier to correctly classify all emails in a test set? Explain why or why not.

No, there will be both false positive results (classifying ham email as spam) and false negative results (classifying spam email as ham). Certain combinations of words may coincidentally occur more (or less) often in the spam training data than the ham training data, and these may cause a test email to be misclassified.

(d) If you were a spammer and you knew we used Naive Bayes to filter spam, how would you change your emails to try to get past the filter?

You could add a lot of words that tend to occur mostly in ham, causing the classifier to misclassify your email as ham. You could purposely misspell spam keywords so that they don't look like words found in the training data.