10-601 Machine Learning: Homework Assignment 2 Solutions

Professor Tom Mitchell
Carnegie Mellon University
January 21, 2009

- The assignment is due at 1:30pm (beginning of class) on **Monday, February 2, 2009**.
- Submit writeups to Problem 1 and Problem 2 *separately* with your name on each problem. Please do not staple the two writeups together.
- Write your name at the top right-hand corner of each page submitted.
- Each student must hand in their own answers to the following questions. See the course webpage for the collaboration policies.
- Each question has the name of the TA who is the primary contact point for that question. Feel free to ask the other instructors about any question, but that TA is the authority on that question.

1 Probability [Purna: 30 points]

1.1 Basic Probability

Consider two events A and B.

1. Use only axioms of probability to prove that $P(A\cap \sim B) = P(A) - P(A\cap B)$
$\star$ *Solution*: We know that $A = (A\cap B)\cup(A\cap \sim B)$. Also the events $A\cap B$ and $A\cap \sim B$ are disjoint events. Hence using $P(A\cup B) = P(A)+P(B)-P(A\cap B)$, we have $P(A) = P(A\cap B)+P(A\cap \sim B)$. Rearranging this gives us the desired result.

2. $P(A\cap B) \geq P(A)+P(B)-1$. This is also known as Bonferroni's Inequality.
$\star$ *Solution*: $P(A\cap B) = P(A)+P(B)-P(A\cup B)$. We know that $P(A\cup B)\leq 1$. Combining these two we get the result.

3. The events A and B are disjoint, if $P(A\cap B) = 0$. If $P(A) = \frac{1}{3}$ and $P(B) = \frac{5}{6}$, then can A and B be disjoint? Explain.
$\star$ *Solution*: We have $P(A\cap B) = P(A) - P(A\cap \sim B)$. Now $A\cap \sim B \subseteq \sim B$, hence $P(A\cap \sim B) \leq P(\sim B) = 1-P(B)$. Now we have $P(A\cap B) \geq P(A)-(1-P(B)) = P(A)+P(B)-1 = \frac{1}{3}+\frac{5}{6}-1 = \frac{1}{6}$. This means that $P(A\cap B)\geq \frac{1}{6}$. In other words $P(A\cap B)$ is strictly greater zero. This implies $A\cap B$ is not empty. Hence A and B are not disjoint.

1.2 Statistical Independence

Two events A and B are statistically independent if $P(A\cap B) = P(A)P(B)$.

1. If A and B are independent events, prove the following
   (a) A and $\sim B$ are independent.
   $\star$ *Solution*: $P(A\cap \sim B) = P(A)-P(A\cap B) = P(A)-P(A)P(B) = P(A)(1-P(B)) = P(A)P(\sim B)$. Hence A and $\sim B$ are also statistically independent.

(b) $\sim A$ and $\sim B$ are independent.
$\star$ *Solution*: $P(\sim A \cap \sim B) = P(\sim B) - P(A \cap \sim B)$. We have already proven that if A and B are independent then A and $\sim B$ are also independent. Hence we have $P(\sim A \cap \sim B) = P(\sim B) - P(A)P(\sim B) = P(\sim B)P(\sim A)$. Hence proved.

2. Rob and Alice are alternately and independently flipping a coin. The first player to get a head wins. Alice flips the coin first.

(a) If $P(head) = \frac{1}{2}$ what is the probability that Alice wins? *hint: Try to enumerate the different settings under which Alice can win!*
$\star$ *Solution*: Note that Alice gets to flip a coin in the odd-numbered runs, i.e. $1, 3, 5, ..$ etc. Now we have,

$$\begin{aligned} Pr(\text{Alice wins}) &= P(\text{Alice wins in turn 1}) + P(\text{Alice wins in turn 3}) + P(\text{Alice wins in turn 5})... \\ &= \sum_{i=0}^{\infty} P(\text{Alice wins in turn } 2 * i + 1) \end{aligned}$$

Now lets compute the probability that Alice wins in turn $2 * i + 1$. This can only happen if no one won until the run $2 * i$, i.e. everyone got tails in all the $2 * i$ turns. The probability of this is $\frac{1}{2}^{2i}$. The probability that Alice will have head in any run is $\frac{1}{2}$. Hence $P(\text{Alice wins in turn } 2 * i + 1) = \frac{1}{2}^{2i+1}$. Hence we want to compute the series sum $\sum_{i=0}^{\infty} \frac{1}{2}^{2i+1} = \frac{1}{2}\sum_{i=0}^{\infty} \frac{1}{4}^{i} = \frac{1}{2}\frac{1}{1-1/4} = \frac{1}{2}\frac{4}{3} = \frac{2}{3}$.

(b) Extra Credit: If $P(head) = p$, then what is the probability that Alice wins? Give your answer in terms of p. *hint: For* $0 \leq a \leq 1$ $\sum_{i=0}^{\infty} a^i = 1/(1-a)$. Given the expression you have derived, would you flip first or second if you were playing the game? Why?
$\star$ *Solution*: Using the exact idea from before, we have $Pr(\text{Alice wins}) = \sum_{i=0}^{\infty}(1-p)^{2i} * p = p\sum_{i=0}^{\infty}((1-p)^2)^i = \frac{p}{1-(1-p)^2}$. Now note that $\frac{p}{1-(1-p)^2} = \frac{p}{p*(2-p)} = \frac{1}{2-p}$. For $p \geq 0$, this is always going to be bigger than 0.5. This means that the probability of winning is greater than random is one starts first. Hence the right strategy will be to flip the coin first.

1.3 Random Variables: Covariance vs. Independence

A random variable is a function mapping the sample space of a random process to real numbers.

1. The **covariance** of two random variables X and Y is defined as

$$Cov(X, Y) = E[(X - E(X))(Y - E(Y))]$$

where $E(X)$ is the expectation of X, and for a discrete X (i.e. X can take discrete values in $\mathcal{X}$) is defined by $\sum_{x \in \mathcal{X}} xP(X = x)$. Prove that

$$Cov(X, Y) = E(XY) - E(X)E(Y)$$

$\star$ *Solution*:

$$\begin{aligned} Cov(X, Y) &= E[(X - E(X))(Y - E(Y))] \\ &= E[XY - XE(Y) - YE(X) + E(X)E(Y)] \\ &= E[XY - XE(Y) - YE(X) + E(X)E(Y)] \\ &= E[XY] - E(X)E(Y) - E(Y)E(X) + E(X)E(Y) \\ &= E(XY) - E(X)E(Y) \end{aligned}$$

We used linearity of expectation, and the fact that $E(C) = C$, is C is a constant.

2. Let X and Y be discrete random variables which take values in $\{0, 1, 2\}$. If you believe the following claims, give a proof, and if not a counter example, i.e. construct a joint probability distribution which disproves the claim.

   (a) If X and Y are independent, their covariance is zero.
   $\star$ *Solution*: Let $\mathcal{X}$ be the set of values the variable X can take.

$$\begin{aligned} E(XY) &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} xyP(X = x, Y = y) \\ &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} xyP(X = x)P(Y = y) \\ &= \sum_{x\in\mathcal{X}} xP(X = x)\sum_{y\in\mathcal{Y}} yP(Y = y) \\ &= E(X)E(Y) \end{aligned}$$

   (b) The converse is also true.
   $\star$ *Solution*: This is false. The example distribution for showing this is:

X	Y	*Probability*
0	0	1/3
1	0	0
2	0	1/3
0	1	0
1	1	1/3
2	1	0

We have $P(X = 0) = 1/3 = P(X = 1) = P(X = 2)$, and $P(Y = 0) = 2/3$, $P(Y = 1) = 1/3$ and $P(Y = 2) = 0$. Here $E(X) = 1$, and $E(Y) = 1/3$. $E(XY) = 1/3$. Hence $cov(X, Y) = 0$. However these aren't independent. $P(X = 0, Y = 0) = 1/3$, whereas $P(X = 0)P(Y = 0) = 2/9$.

1.4 Conditional Probabilities

By now you all know the definition of conditional probability. It is defined as

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{1}$$

In this question we will see how the probability of an event can change given our knowledge about another related event. Two fair die are rolled together. Let the random variable S denote the sum of the numbers read from the two.

1. What is the probability that $S = 11$?
   $\star$ *Solution*: There are 36 different values S can take. Lets look at number of cases when $S = 11$. This is the set $\{< 5, 6 >, < 6, 5 >\}$. Hence $P(S = 11) = 2/36 = 1/18$.

2. If you know that the S is a prime number, then what is the above probability?
   $\star$ *Solution*: Let us denote by M the event which consists of those cases where S is prime. $\{< 1, 1 >, < 1, 2 >, < 1, 4 >, < 1, 6 >, < 2, 1 >, < 2, 3 >, < 2, 5 >, < 3, 2 >, < 3, 4 >, < 4, 1 >$

$, < 4, 3 >, < 5, 2 >, < 5, 6 >, < 6, 1 >, < 6, 5 >\}$. We want $P(S = 11|M)$. Which is simply, $P(S = 11|M) = \frac{P(S=11 \cap M)}{P(M)} = \frac{P(S=11)}{P(M)}$. We can plug in the values to get $\frac{2/36}{15/36} = \frac{2}{15}$. Note that the event $S = 11$ is a subset of M, which is why we have $P(S = 11 \cap M) = P(S = 11)$.

2 Conditional independence and parameter estimation [Andy: 30 points]

Note: Throughout this problem, when asked about estimates, we are concerned with MLEs and not MAP estimates. Assume you are given a training dataset comprised of m = 1,000 binary classed examples (500 in the positive $y = 1$ class, 500 in the negative $y = 0$ class), each consisting of n = 10 binary valued attributes, generated from the following model, M_{indp}, which assumes conditional independence between attributes, given their class:

$$\mathrm{M}_{\mathrm{indp}} : \forall i : 1 \leq i \leq n, \forall x, y \in \{0, 1\} \quad Pr(X_i = x|Y = y) = p^{indp}_{i,x,y}$$

In other words, each example $\langle\langle x_1, x_2, \ldots, x_n\rangle, y\rangle$ is generated by first picking a value y for the class Y, then picking the value x_i of each attribute X_i with probability $p^{indp}_{i,x_i,y}$. Each attribute x_i is thus determined independently of the other attributes. We will also assume that the probability of picking class $Y = 1$ is 0.5, i.e. $P(Y = 1) = P(Y = 0) = 0.5$.

1. How many free parameters, $p^{indp}_{i,x,y}$, does this model have?

   ⋆ *Solution*:

   The distribution of each $Pr(X_i = x|Y = y)$ for some specific i and y has one free parameter. i has 10 different values and Y has 2, so there are 20 free parameters.

2. Now assume you are given a particular instance of such a model, where the parameters are set as follows: $\forall i : p^{indp}_{i,1,1} = .8$ and $p^{indp}_{i,1,0} = .6$ (i.e., the probability of any attribute being set to 1 is 0.8 for a positive example, and 0.6 for a negative example). Assume you are also given a single test example from the **positive** class, $\langle \bar{x}_{test}, y_{test}\rangle = \langle\langle 1, 1, 0, 0, 1, 1, 0, 1, 1, 1\rangle, 1\rangle$.

   What is the probability of the instance $\bar{x}_{test} = \langle 1, 1, 0, 0, 1, 1, 0, 1, 1, 1\rangle$ being generated given that the class is positive—in other words, what is $Pr(\bar{x}_{test}|y = 1, M_{indp})$?

   ⋆ *Solution*:

$$\begin{aligned} & Pr(\bar{x}_{test}|y = 1, M_{indp}) \\ = \; & p^{indp}_{1,1,1} \cdot p^{indp}_{2,1,1} \cdot p^{indp}_{3,0,1} \cdot p^{indp}_{4,0,1} \cdot p^{indp}_{5,1,1} \cdot p^{indp}_{6,1,1} \cdot p^{indp}_{7,0,1} \cdot p^{indp}_{8,1,1} \cdot p^{indp}_{9,1,1} \cdot p^{indp}_{10,1,1} \\ = \; & 0.8^7 \cdot 0.2^3 \end{aligned}$$

3. What is the $Pr(y = 1|\bar{x}_{test}, M_{indp})$? (I.e., what is the predicted probability that the class is 1 **under the model** defined in Part 2?)

   ⋆ *Solution*:

   Using Bayes' rule:

$$\begin{aligned} Pr(y = 1|\bar{x}_{test}, M_{indp}) &= \frac{Pr(\bar{x}_{test}|y = 1)Pr(y = 1)}{Pr(\bar{x}_{test}|y = 0)Pr(y = 0) + Pr(\bar{x}_{test}|y = 1)Pr(y = 1)} \\ &= \frac{0.8^7 \cdot 0.2^3 \cdot 0.5}{0.8^7 \cdot 0.2^3 \cdot 0.5 + 0.6^7 \cdot 0.4^3 \cdot 0.5} \\ &\approx 0.486 \end{aligned}$$

4. Based on the training data, what is the maximum likelihood estimator $\hat{p}^{indp}_{i,x,1}$ for the model parameter $p^{indp}_{i,x,1}$? What is the MLE $\hat{p}^{indp}_{i,x,0}$ for the model parameter $p^{indp}_{i,x,0}$? Express your answer in terms of *properties of the training data*, not the instantiated model parameters given in (2) above).

   $\star$ *Solution*:

   Using the notation from class, where $\#D\{\cdot\}$ indicates the number of training data points with some particular property:

$$\hat{p}^{indp}_{i,x,y} = \frac{\#D\{X_i = x \wedge Y = y\}}{\#D\{Y = y\}}$$

5. Now consider a new model, M_{dep}, where *no* assumptions are made regarding the possible dependencies between attributes:

$$\mathrm{M_{dep}} : \forall \bar{x} : \bar{x} \in \{0,1\}^n, \forall y \in \{0,1\} \quad Pr(\bar{X} = \bar{x} | Y = y) = p^{dep}_{\bar{x},y}$$

   In other words, each example $\langle\langle x_1, x_2, \ldots, x_n\rangle, y\rangle$ is generated by first picking a value y for the class Y, then picking an entire vector $\bar{x} = \langle x_1, x_2, \ldots, x_n\rangle$, with the probability of picking that vector given by the parameter $p^{dep}_{\bar{x},y}$. We will still assume that the probability of picking class $Y = 1$ is 0.5, i.e. $P(Y = 1) = P(Y = 0) = 0.5$.

   How many free parameters, $p^{dep}_{\bar{x},y}$, does this model have? How does this compare to M_{indp}?

   $\star$ *Solution*:

   The distribution of each $Pr(\bar{X} = \bar{x}|Y = y)$ for some specific value of y has $2^n - 1$ free parameters, since $\bar{X}$ has 2^n unique values, and $\sum_{\bar{x}} Pr(\bar{X} = \bar{x}|Y = y)$ must be equal to 1. Y has 2 different values, so there are $2(2^n - 1)$ free parameters. This is exponential in n, compared to a number of free parameters linear in n for M_{indp}.

6. Let $\bar{x}_{test}$ refer to the single test example of Part (2). Under this new M_{dep}, to find $Pr(Y|\bar{x}_{test})$ you first need to estimate $p^{dep}_{\bar{x}_{test},1}$ and $p^{dep}_{\bar{x}_{test},0}$. Given that you have 500 training examples of each class generated from $M_{\mathbf{indp}}$, but learned your estimates $\hat{p}^{dep}_{\bar{x}_{test},1}$ and $\hat{p}^{dep}_{\bar{x}_{test},0}$ over this training data assuming $M_{\mathbf{dep}}$:

   (a) What is the MLE $\hat{p}^{dep}_{\bar{x}_{test},1}$ for the parameter $p^{dep}_{\bar{x}_{test},1}$? (again, express this in terms of properties of the training data.)

   $\star$ *Solution*:

$$\hat{p}^{dep}_{\bar{x}_{test},1} = \frac{\#D\{\bar{X} = \bar{x}_{test} \wedge Y = 1\}}{\#D\{Y = 1\}}$$

   (b) Given that the training data was generated from M_{indp} using the parameters given in Part 2, what is the probability that this MLE will be zero? I.e., what is $Pr(\hat{p}^{dep}_{\bar{x}_{test},1} = 0)$, where the probability here is taken over different outcomes of the "experiment" of generating the training data from M_{indp}.

   $\star$ *Solution*: The probability of a given positive training example NOT being equal to $\bar{x}_{test}$ is $1 - Pr(\bar{x}_{test}|y = 1, M_{indp})$. The probability of this happening 500 times in a row is $(1 - Pr(\bar{x}_{test}|y = 1, M_{indp}))^{500} = (1 - 0.8^7 \cdot 0.2^3)^{500} \approx 0.432$.

(c) What is $Pr(\hat{p}^{dep}_{\bar{x}_{test},0} = 0)$, assuming again that the data was generated using the M_{indp} model from Part 2?

⋆ *Solution*: The probability of a given negative training example NOT being equal to $\bar{x}_{test}$ is $1 - Pr(\bar{x}_{test}|y = 0, M_{indp})$. The probability of this happening 500 times in a row is $(1 - Pr(\bar{x}_{test}|y = 0, M_{indp}))^{500} = (1 - 0.6^7 \cdot 0.4^3)^{500} \approx 0.408$.

7. Consider this new set of training data:

X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	Y
0	1	0	1	1	1	0	1	1	0	0
1	1	0	1	1	1	0	1	1	0	0
1	1	1	0	1	1	1	0	1	0	0
1	1	1	1	0	1	0	0	1	0	1
0	1	0	0	0	1	1	1	0	0	1
1	1	0	0	0	1	1	0	0	0	1

(a) Based on this new training data, what are the maximum likelihood estimates $\hat{p}^{indp}_{i,x,y}$ for the parameters of the model M_{indp}?

⋆ *Solution*:

The MLE estimates are:

i	1	2	3	4	5	6	7	8	9	10
$\hat{p}^{indp}_{i,1,1}$	2/3	1	1/3	1/3	0	1	2/3	1/3	1/3	0
$\hat{p}^{indp}_{i,1,0}$	2/3	1	1/3	2/3	1	1	1/3	2/3	1	0

(b) As we discussed in class, Dirichlet priors are commonly used when estimating parameters to avoid zeros. If we assume a Dirichlet prior over each of the parameters in M_{indp} where the parameters to the Dirichlet are $\alpha_0 = \alpha_1 = 2$, what are the MAP estimates for those same $p^{indp}_{i,x,y}$?

⋆ *Solution*:

The MAP estimates are:

i	1	2	3	4	5	6	7	8	9	10
$\hat{p}^{indp}_{i,1,1}$	3/5	4/5	2/5	2/5	1/5	4/5	3/5	2/5	2/5	1/5
$\hat{p}^{indp}_{i,1,0}$	3/5	4/5	2/5	3/5	4/5	4/5	2/5	3/5	4/5	1/5

8. In one or two sentences, how does this problem relate to the discussion in class about conditional independence and Naïve Bayes?

⋆ *Solution*: M_{indp} corresponds to the naive Bayes model, while M_{dep} corresponds to a model that makes no conditional independence assumptions.