# 10-601 Machine Learning: Homework Assignment 2

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 $P(A \cap \sim B) = P(A) - P(A \cap B)$
2. $P(A \cap B) \geq P(A) + P(B) - 1$. This is also known as Bonferroni's Inequality.
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.

### 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.
   (b) $\sim A$ and $\sim B$ are independent.
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!*
   (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?

### 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

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

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.

   (b) The converse is also true.

### 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 $S = 11$?

2. If you know that the $S$ is a prime number, then what is the above probability?

## 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:

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

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?

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 $Pr(\bar{x}_{test}|y = 1, M_{indp})$?

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?)

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).

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 $M_{indp}$?

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.)

(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 $M_{indp}$.

(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?

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 $M_{indp}$?

(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 $\alpha_0 = \alpha_1 = 1$, what are the MAP estimates for those same $p^{indp}_{i,x,y}$?

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