Solution for Assignment 1 :
Intro to Probability and Statistics, PAC learning

10-701/15-781: Machine Learning (Fall 2004)

Due: Sept. 30th 2004, Thursday, Start of class

Question 1. Basic Probability (18 pts)

- 1.1 (2 pts) Suppose that A is an event such that $Pr(A) = 0$ and that B is any other event. Prove that A and B are independent events.
  Answer:
  Since the event $A \wedge B$ is a subset of the event A, and Pr(A) = 0, so $Pr(A \wedge B) = 0$. Hence $Pr(A \wedge B) = 0 = Pr(A) * Pr(B)$.

- 1.2 (3 pts) Prove: Let $a_1, a_2, ..., a_n$ be possible values of A. Then for any event B, $P(B = b) = \sum_{i=1}^{n} P(B = b|A = a_i) * P(A = a_i)$

  Answer:
  $P(B = b) = P(B = b \wedge (TRUE))$
  $= P(B = b \wedge (A = a_1 \vee A = a_2 \vee ... \vee A = a_n))$
  $= P((B = b \wedge A = a_1) \vee (B = b \wedge A = a_2)... \vee (B = b \wedge A = a_n)))$
  $= P(B = b \wedge A = a_1) + P((B = b \wedge A = a_2) \vee ... \vee (B = b \wedge A = a_n))) - P((B = b \wedge A = a_1) \wedge ((B = b \wedge A = a_2) \vee ... \vee (B = b \wedge A = a_n))))$
  $= P(B = b \wedge A = a_1) + P((B = b \wedge A = a_2) \vee ... \vee (B = b \wedge A = a_n)))$
  $= ...$
  $= \sum_{i=1}^{n} P(B = b \cap A = a_i)$
  $= \sum_{i=1}^{n} P(B = b|A = a_i) * P(A = a_i)$

- 1.3 (5 pts) Soldier A and Soldier B are practicing shooting. The probability that A would miss the target is 0.2 and the probability that B would miss the target is 0.5. The probability that both A and B would miss the targets is 0.1.

  - What is the probability that at least one of the two will miss the target?
  Answer: $P(A \vee B) = P(A) + P(B) - P(A \wedge B) = 0.6$

  - What is the probability that exactly one of the two soldiers will miss the target?
  Answer: $P(A \vee \bar{B}) + P(B \vee \bar{A}) = 0.5$

- 1.4 (4 pts) A box contains three cards. One card is red on both sides, one card is green on both sides, and one card is red on one side and green on the other. Then we randomly select one card from this box, and we can know the color of the selected card's upper side. If this side is green, what is the probability that the other side of the card is also green?
  Answer:
  $P(the - other - side - green|this - side - green) = \frac{P(both-sides-green)}{P(this-side-green)} = \frac{1/3}{1/2} = \frac{2}{3}$

- 1.5 (4 pts) Suppose that the p.d.f. of a random variable X is:

$$f(x) = \begin{cases} cx^2, & for 1 \leq x \leq 2 \\ 0, & otherwise \end{cases} \tag{1}$$

- What is the value of constant c?
Answer:
$\int_1^2 cx^2 dx = \frac{7}{3}c = 1 \rightarrow c = \frac{3}{7}$

- Sketch the p.d.f.
- $Pr(X > 3/2) =?$
Answer: $\int_{3/2}^2 cx^2 dx = 37/56$

Question 2. Expectation (18 pts)

- 2.1 (4 pts) If an integer between 100 and 200 is to be chosen at random, what is the expected value?
Answer: $E(X) = \frac{1}{101}(100 + 101 + ... + 200) = 150$
- 2.2 (5 pts) A rabbit is playing a jumping game with friends. She starts from the origin of a real line and moves along the line in jumps of one step. For each jump, she flips a coin. If heads, she would jump one step to the left (i.e. negative direction). Otherwise, she would jump one step to the right. The chance of heads is p ($0 \leq p \leq 1$). What is the expected value of her position after n jumps ? (assume each step is in equal length and assume one step as one unit on the real line)
Answer:
For the ith jumping, $E(X_i) = (-1)p + (1)(1-p) = 1 - 2p$, So the position after n jumps is:
$E(X_1 + X_2 + ... + X_n) = E(X_1) + E(X_2) + ... + E(X_n) = n(1-2p)$
- 2.3 (4 pts) Suppose that the random variable X has a uniform distribution on interval [0, 1]. Random variable Y has a uniform distribution on the interval [4, 10]. X and Y are independent. Suppose a rectangle is to be constructed for which the lengths of two adjacent sides are X and Y. So what is the expected value of the area of this rectangle?
Answer:
Since X and Y are independent, $E(X \wedge Y) = E(X) * E(Y) = 0.5 * 7 = 3.5$
- 2.4 (5 pts) Suppose that X is a random variable. $E(X) = \mu, Var(X) = \sigma^2$, then what is the value of $E[X(X-1)] =?$
Answer:
$E[X(X-1)] = E[X^2] - E[X] = var(X) + E[X]^2 - \mu = \sigma^2 + \mu^2 - \mu$

Question 3. Normal Distribution (6 pts)

Suppose X has a normal distribution with mean 1 and variance 4. Find the value of the following:
a. $Pr(X \leq 3)$
Answer:
$Pr(X \leq 3) = Pr(Z \leq \frac{3-1}{\sqrt{4}}) = \Phi(1) = 0.8413$

b. $Pr(|X| \leq 2)$
Answer: $Pr(|X| \leq 2) = \Phi(\frac{2-1}{2}) - \Phi(\frac{-2-1}{2}) = \Phi(0.5) - (1 - \Phi(1.5)) = 0.6247$

Question 4. Bayes' Theorem (8 pts)

In a certain day care class, 30 percent of the children have grey eyes, 50 percent of the children have blue eyes, and the other 20 percent's eyes are in other colors. One day they play a game together. In the first run, 65 percent of the grey eye kids were selected into the game, 82 percent of the blue eye kids selected in, and 50 percent of the kids with other colors were chosen. So if a child is selected at random from the class, and we know that he was not in the first run game, what is the probability that he has blue eyes?
Answer:
Assume B: blue eyes; O: other color eyes; G: grey eyes; NF: not in the first run game
$P(B|NF) = \frac{P(B)P(NF|B)}{P(B)P(NF|B)+P(O)P(NF|O)+P(G)P(NF|G)} = \frac{0.5*0.18}{0.5*0.18+0.2*0.5+0.3*0.35} = 0.3051$

Question 5. Probabilistic Inference (15 pts)

Imagine there are three boxes labelled A, B and C. Two of them are empty, and one contains a prize. Unfortunately, they are all closed and you don't know where the prize is. You first pick a box at random, say box A. However, before you open it, box B is opened by someone, and you see that it is empty. You now have to make your final choice as to what box to open: A or C.

Question: For each of the cases below, answer what box would you open so as to maximize the chances that the box you open contains the prize. Support your arguments by computing the probability of the prize being in box A and C.

Here are the three strategies according to which box B was chosen to be opened:

1. (5 pts) In this strategy if you first pick a box (in this case A) with a prize, then one of the other two boxes is opened at random. On the other hand, if you first choose a box that has no prize, then the empty box that you did not pick is chosen.

2. (5 pts) In this strategy it is just one of the two boxes that you did not pick is chosen at random (in this case it is a random choice between B and C).

3. (5 pts) In this strategy one of empty boxes is chosen at random (independently of whether you initially pick a box with a prize or not).

Answer:
Let SpB stand for a random event of "someone picks box B". In all the cases the prior (before box B was opened) probabilities that prize is in box A, B, or C are $P(A) = P(B) = P(C) = 1/3$. The differences are in the conditional probabilities: $P(SpB|A), P(SpB|B), P(SpB|C)$. In all three cases we compute posterior (after box B was opened) probabilities. We then pick a box with the highest probability of containing a prize.

1. $P(SpB|A) = 1/2$; $P(SpB|B) = 0$; $P(SpB|C) = 1$;
   $P(A|SpB) = \frac{P(SpB|A)P(A)}{P(SpB)}$ and $P(C|SpB) = \frac{P(SpB|C)P(C)}{P(SpB)}$;
   $P(SpB) = P(SpB|A)P(A) + P(SpB|B)P(B) + P(SpB|C)P(C) = 1/2 * 1/3 + 0 * 1/3 + 1 * 1/3 = 1/2$;
   $P(A|SpB) = \frac{1/2*1/3}{1/2} = 1/3$ and $P(C|SpB) = \frac{1*1/3}{1/2} = 2/3$;

2. $P(SpB|A) = 1/2$; $P(SpB|B) = 1/2$; $P(SpB|C) = 1/2$;
   Unlike in the previous sub-question, here the box that was opened by someone (namely, box B) could have contained a prize. Therefore, the posterior probabilities we are interested in are: $P(A|SpB \wedge \neg B)$ and $P(C|SpB \wedge \neg B)$.
   $P(A|SpB \wedge \neg B) = \frac{P(SpB\wedge\neg B|A)P(A)}{P(SpB\wedge\neg B)}$ and $P(C|SpB \wedge \neg B) = \frac{P(SpB\wedge\neg B|C)P(C)}{P(SpB\wedge\neg B)}$;
   $P(SpB \wedge \neg B|A) = P(SpB|A)P(\neg B|A) = 1/2 * 1 = 1/2$
   $P(SpB \wedge \neg B|B) = 0$;
   $P(SpB \wedge \neg B|C) = P(SpB|C)P(\neg B|C) = 1/2 * 1 = 1/2$
   $P(SpB \wedge \neg B) = P(SpB \wedge \neg B|A)P(A) + P(SpB \wedge \neg B|B)P(B) + P(SpB \wedge \neg B|C)P(C) = 1/2 * 1/3 + 0 * 1/3 + 1/2 * 1/3 = 1/3$;
   $P(A|SpB \wedge \neg B) = \frac{1/2*1/3}{1/3} = 1/2$ and $P(C|SpB \wedge \neg B) = \frac{1/2*1/3}{1/3} = 1/2$;

3. $P(SpB|A) = 1/2$; $P(SpB|B) = 0$; $P(SpB|C) = 1/2$;
   $P(A|SpB) = \frac{P(SpB|A)P(A)}{P(SpB)}$ and $P(C|SpB) = \frac{P(SpB|C)P(C)}{P(SpB)}$;
   $P(SpB) = P(SpB|A)P(A) + P(SpB|B)P(B) + P(SpB|C)P(C) = 1/2*1/3 + 0*1/3 + 1/2*1/3 = 1/3$;
   $P(A|SpB) = \frac{1/2*1/3}{1/3} = 1/2$ and $P(C|SpB) = \frac{1/2*1/3}{1/3} = 1/2$;

Question 6. PAC-learning I (15pts)

Consider an image classification problem. Suppose an algorithm first splits each image into $n = 4$ blocks (the blocks are non-overlapping and each block is at the same location and of constant size across all images) and computes some scalar feature value for each of the blocks (e.g., average intensity of the pixels within the block). Suppose that this feature is discrete and can take $m = 10$ values. The classification function classifies an image as 1 whenever each of the n feature values lies within some interval that is specific to this feature (i.e., the value of the first feature is between a_1 and b_1, the value of the second feature is between a_2 and b_2, and so on), and 0 otherwise. We would like to learn these intervals (a and b values for each interval) automatically based on a training set of images. All the other parameters such as locations and sizes of the blocks are *not* being learned. The following questions are helpful in understanding the requirements on the size of the training set.

1. (7 pts) What is the size of the hypothesis space H? Assume that only intervals with $a_i \leq b_i$ are considered for learning.
2. (4 pts) Assuming noiseless data and that the function we are trying to learn is capable of perfect classification, give an upper bound on the size of the training set required to be sure with 99% probability that the learned function will have true error rate of at most 5%.
3. (4 pts) Compare $|H|$ (the answer to question 1) and the required training dataset size R (the answer to question 2). Why does R not seem to be very affected by the number of possible hypotheses? What parameter does make R increase quickly and why? (Please provide only a few sentences for each question).

Answer:

1. The number of possible intervals for any particular feature value is computed as follows: for a given a_i the possible b_i values are from a_i to m (that is, $m - a_i + 1$ values); hence, the number of possible intervals is $m + (m-1) + (m-2) + \ldots + 1 = \frac{m(m+1)}{2}$;
   Since there are n features and we use a boolean conjunction function $|H| = (\frac{m(m+1)}{2})^n = 55^4$;
2. $R \geq \frac{0.69}{\epsilon}(log_2(55^4) + log_2(1/\delta)) = 410.82$;
3. In terms of formula, R is logarithmically related to the number of possible hypotheses and inverse proportionally to ϵ. Thus, ϵ affects R much stronger. Intuitively speaking, if a learned hypothesis is consistent with a large number of iid data points, then chances are it will classify correctly the test data points as well. This will hold independently of how many hypotheses we have. On the other hand, in order to guarantee a very small misclassification rate (ϵ), the hypothesis needs to be trained on a very large number of samples (so that they cover almost all of the input space).

Question 7. PAC-learning II (20pts)

Consider a learning problem in which input datapoints are real numbers distributed uniformly in between a and b, and output is binary. The true function we are trying to learn is $x < c^*$ for some $a \leq c^* \leq b$ (that is, output 1 whenever $x < c^*$ and 0 otherwise). The set of hypotheses is therefore: $H = \{(x < c)|a \leq c \leq b\}$ (the hypothesis space is therefore infinite: all real values of c in between a and b). Assuming that we have m datapoints for training, derive an upper bound on the probability of learning a hypothesis that will have

a true classification error larger than ϵ. The derivation should be done in the same spirit as the one used to derive a PAC bound on the probability of learning a 'bad' h for the case when hypothesis space H is finite. Do not use bounds that are based on VC-dim (please ignore this sentence if you do not know VC-dim anyway). Give the bound in terms of a, b, c^*, m and ϵ. Then evaluate it numerically for the following values: $a = 0$, $b = 2$, $c^* = 1$, $m = 20$ and $\epsilon = 0.1$. (Hint: you may need to use integrals).

Answer:

There are few possible answers to this. Here are some (others are also possible):

1.

$$
\begin{aligned}
P(\text{we learn h' such that } \mathit{trueerror}(h') > \epsilon) &\leq \\
P(\text{the set H contains h' such that } \mathit{trueerror}(h') > \epsilon) &= \\
P(\exists h',\ h' \text{ is consistent with } m \text{ examples } \mathit{and}\ \mathit{trueerror}(h') > \epsilon) &= \\
P(\exists c, c \text{ is consistent with } x_1 \ldots x_m\ \mathit{and}\ \mathit{trueerror}(h = (x < c)) > \epsilon) &\leq \\
P(x_1 \ldots x_m \notin [\max(c^* - \epsilon(b-a), a); c^*]\ \mathit{or}\ x_1 \ldots x_m \notin [c^*; \min(c^* + \epsilon(b-a), b)]) &= \\
P(x_1 \ldots x_m \notin [\max(c^* - \epsilon(b-a), a); c^*]) + & \\
P(x_1 \ldots x_m \notin [c^*; \min(c^* + \epsilon(b-a), b)]) - & \\
P(x_1 \ldots x_m \notin [\max(c^* - \epsilon(b-a), a); c^*]\ \mathit{and}\ x_1 \ldots x_m \notin [c^*; \min(c^* + \epsilon(b-a), b)]) &= \\
(1 - \frac{c^* - \max(c^* - \epsilon(b-a), a)}{b-a})^m + & \\
(1 - \frac{\min(c^* + \epsilon(b-a), b) - c^*}{b-a})^m - & \\
(1 - \frac{\min(c^* + \epsilon(b-a), b) - \max(c^* - \epsilon(b-a), a)}{b-a})^m &
\end{aligned}
$$

For a well-behaved c^* we have: $\delta \leq 2(1-\epsilon)^m - (1-2\epsilon)^m$.

2. This version is *almost* correct and deserves a full credit if given.

$$
\begin{aligned}
P(\text{we learn h' such that } \mathit{trueerror}(h') > \epsilon) &\leq \\
P(\text{the set H contains h' such that } \mathit{trueerror}(h') > \epsilon) &= \\
P(\exists h',\ h' \text{ is consistent with } m \text{ examples } \mathit{and}\ \mathit{trueerror}(h') > \epsilon) &= \\
P(\exists c, c \text{ is consistent with } x_1 \ldots x_m\ \mathit{and}\ \mathit{trueerror}(h = (x < c)) > \epsilon) &\leq \\
\int_a^b P(c \text{ is consistent with } x_1 \ldots x_m\ \mathit{and}\ \mathit{trueerror}(h = (x < c)) > \epsilon) dc &= \\
\int_a^b P(x_1 \ldots x_m \notin [c, c^*] \text{ if } c < c^* \text{ or } x_1 \ldots x_m \notin [c^*, c] \text{ if } c > c^*\ \mathit{and}\ |c, c^*| > \epsilon(b-a)) dc &= \\
\int_a^{\max(c^* - \epsilon(b-a), a)} P(x_1 \ldots x_m \notin [c, c^*]) dc + \int_{\min(c^* + \epsilon(b-a), b)}^b P(x_1 \ldots x_m \notin [c^*, c]) dc &= \\
\int_a^{\max(c^* - \epsilon(b-a), a)} (1 - (c^* - c)/(b-a))^m dc + \int_{\min(c^* + \epsilon(b-a), b)}^b (1 - (c - c^*)/(b-a))^m dc &= \\
1/(b-a)^m * (\int_a^{\max(c^* - \epsilon(b-a), a)} (b - a - c^* + c)^m dc + \int_{\min(c^* + \epsilon(b-a), b)}^b (b - a - c + c^*)^m dc) &= \\
1/((m+1)(b-a))^m * ((b - a - c^* + \max(c^* - \epsilon(b-a), a))^{m+1} - (b - a - c^* + a)^{m+1} - & \\
(b - a - b + c^*)^{m+1} + (b - a - \min(c^* + \epsilon(b-a), b) + c^*)^{m+1}) &= \\
1/((m+1)(b-a))^m * ((b - a - c^* + \max(c^* - \epsilon(b-a), a))^{m+1} - (b - c^*)^{m+1} - & \\
(-a + c^*)^{m+1} + (b - a - \min(c^* + \epsilon(b-a), b) + c^*)^{m+1}) &
\end{aligned}
$$