Homework 1

Michael Anderson

April 9, 2011

CS534

Prof. Fern

1

1. By definition:

$$E(aX + bY) = \sum_i ax_i P_x(x_i) + by_i P_y(y_i)$$

$$= \sum_i ax_i P_x(x_i) + \sum_i by_i P_y(y_i)$$

Since a and b are constant with respect to i, they can be pulled out of the sum to get:

$$a \sum_i x_i P_x(x_i) + b \sum_i y_i P_y(y_i) = aE(x) + bE(y)$$

This is easily generalized to include continuous random variables, because their expectation formula is simply a sum of infinitesimals.

2.

$$Var(aX + bY) = \sum_i (ax_i - a\bar{x})^2 P_x(x_i) + (by_i - b\bar{y})^2 P_y(y_i)$$

$$= \sum_i a^2 (x_i - \bar{x})^2 P_x(x_i) + \sum_i b^2 (y_i - \bar{y})^2 P_y(y_i)$$

$$= a^2 \sum_i (x_i - \bar{x})^2 P_x(x_i) + b^2 \sum_i (y_i - \bar{y})^2 P_y(y_i)$$

$$= a^2 Var(X) + b^2 Var(Y)$$

3.

$$Cov(X, Y) = E((X - \bar{x})(Y - \bar{y}))$$

$$= \sum_i [(x_i - \bar{x}) P_x(x_i) \times (y_i - \bar{y}) P_y(y_i)]$$

Since it is given that X and Y are independent, the value of i in one does not affect the other, and the sum can be separated into two parts. Each of those parts can also be separated into two parts:

$$= [\sum_i x_i P_x(x_i) - \sum_i \bar{x} P_x(x_i)][\sum_i y_i P_y(y_i) - \sum_i \bar{y} P_y(x_i)]$$

$$= [E(X) - \bar{x}][E(Y) - \bar{y}] = [0][0] = 0$$

2

Calculate the CDF (Cumulative Distribution Function), then differentiate to get the PDF. For some value $0 \leq t \leq 1$:

$$CDF(t) = P(max(X, Y) < t)$$

$$= 1 - (1-t)(1-t) = 1 - (1 - 2t + t^2) = 2t - t^2$$

Now:

$$PDF(t) = \frac{d}{dt}CDF(t) = \frac{d}{dt}(2t - t^2) = -2t + 2$$

3

The probability of grabbing some orange is the product of the probabilities of two independent events. The probability of selecting the box containing an orange, and the probability of selecting an orange from the available fruit in the box. One of the three boxes will be selected with equal probability.

$$P(Orange) = \frac{1}{3} \times \frac{3}{3+6} + \frac{1}{3} \times \frac{3}{3+3} + \frac{1}{3} \times \frac{5}{5+3}$$

$$= \frac{3}{27} + \frac{3}{18} + \frac{5}{24}$$

$$\approx 0.486$$

4

In this example the loss from a false negative l_0 is twice the loss of a false positive l_1, so we would only want to predict 0 if the $p_0 > 2p_1$, else predict 1. Since $p_0 + p_1 = 1$, we have the threshold at $p_0 = 2/3$, $p_1 = 1/3$. I.e. $\theta = 1/3$.

Generally, $\theta = \frac{l_1}{l_0+l_1}$. If for example $\theta = 0.10$ and $l_1 = 1$, then $l_0 = 9$, giving the following loss matrix:

$\hat{y} \backslash y$	0	1
0	0	9
1	1	0

5

Let E_0, E_1, and E_r be the expected loss of predicting 0, predicting 1, and rejecting respectively. Let l_0 be the loss of a false negative, l_1 be the loss of a false positive, and l_r be the loss of rejecting. Then:

$$E_0 = p_1 l_0$$

$$E_1 = (1 - p_1) l_1$$

$$E_r = l_r$$

To figure out which of the three decisions is best for $p_1 = 0.2$, pick the decision which minimizes the expected loss.

$$min(p_1 l_0, (1 - p_1) l_1, l_r)$$

$$= min(0.2 \times 10, 3, 0.8 \times 10) = 2 = E_0$$

So predict 0.

Generally, E_0 increases linearly as p_1 increases, E_r stays constant, and E_1 decreases linearly as p_1 increases. This suggests that there will be some interval starting at $p_1 = 0$ when E_0 is the smallest, then an intermediate interval when E_r is smallest, and in the final interval ending at $p_1 = 1$ where E_1 will be the smallest. To find the endpoints of these intervals, set $E_0 = E_r$ and then $E_r = E_1$.

$$E_0 = E_r \quad \rightarrow \quad p_1 l_0 = l_r \quad \rightarrow \quad p_1 = l_0 / l_r$$

$$E_r = E_1 \quad \rightarrow \quad l_r = (1 - p_1) l_1 \quad \rightarrow \quad p_1 = 1 - \frac{l_r}{l_1}$$

Therefore $\theta_0 = l_0 / l_r$, and $\theta_1 = 1 - l_r / l_1$. Note that if $l_r > l_0$ or $l_r > l_1$, we end up with $\theta_0 < 0$ or $\theta_1 > 1$. This means that at no point in the interval does the reject option minimize loss, and the problem reduces to one like Exercise 4.