Homework 1

10-601 Machine Learning
Name: Shashank Singh
Email: sss1@andrew.cmu.edu
Due: Friday, September 14, 2012

---

1 Probability Review

(a) **Equation of the Reverend**
By definition of conditional probability, for any events A and B,

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \quad \text{and} \quad P(B|A) = \frac{P(A \cap B)}{P(A)}.$$

Multiplying by $P(A)$ in the second equation gives $P(B|A)P(A) = P(A \cap B)$. Thus, substituting into the first equation,

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}. \quad \blacksquare$$

(b) **Contingencies**

1.
$$P(A = \diamondsuit) = \frac{12+3}{12+3+97+5} = \boxed{\frac{15}{117}.}$$

2.
$$P(A = \diamondsuit \text{ AND } B = \square) = \frac{3}{12+3+97+5} = \boxed{\frac{3}{117}.}$$

3.
$$P(A = \diamondsuit \text{ OR } B = \square) = \frac{12+3+5}{12+3+97+5} = \boxed{\frac{20}{117}.}$$

4.
$$P(A = \diamondsuit | B = \square) = \frac{P(A = \diamondsuit \text{ AND } B = \square)}{P(B = \square)} = \frac{\frac{3}{117}}{\frac{3+5}{117}} = \boxed{\frac{3}{8}.}$$

5. By the Law of Total Probability,

$$\begin{aligned}
P(A = \diamondsuit) &= P(A = \diamondsuit | B = \triangle)P(B = \triangle) + P(A = \diamondsuit | B = \square)P(B = \square) \\
&= P(A = \diamondsuit | B = \triangle) \cdot \frac{12+97}{117} + P(A = \diamondsuit | B = \square) \cdot \frac{3+5}{117} \\
&= \frac{12}{12+97} \cdot \frac{12+97}{117} + \frac{3}{3+5} \cdot \frac{3+5}{117} = \boxed{\frac{15}{117}.}
\end{aligned}$$

(c) **Chain Rule**

$$\begin{aligned} P(X,Y,Z) &= P(X,Y|Z)\cdot P(Z) \\ &= P(X|Y,Z)\cdot P(Y|Z)\cdot P(Z) \end{aligned}$$

(d) **Total Probability and Independence**

1.

$$\begin{aligned} P(X=1) &= P(X=1|Y=0)\cdot P(Y=0) + P(X=1|Y=1)\cdot P(Y=1) \\ &= P(X=1|Y=0)\cdot P(Y=0) \\ &+ P(X=1|Y=1,Z=1)\cdot P(Z=1|Y=1)\cdot P(Y=1) \\ &+ P(X=1|Y=1,Z=0)\cdot P(Z=0|Y=1)\cdot P(Y=1) \\ &= P(X=1|Y=0)\cdot P(Y=0) \\ &+ P(X=1|Y=1,Z=1)\cdot P(Z=1)\cdot P(Y=1) \\ &+ P(X=1|Y=1,Z=0)\cdot P(Z=0)\cdot P(Y=1) & \text{since } Z \perp Y \\ &= (0.2)(0.1) + (0.6)(0.8)(0.9) + (0.1)(0.2)(0.9) = \boxed{0.47.} \end{aligned}$$

2.

$$E[Y] = 0\cdot P(Y=0) + 1\cdot P(Y=1) = P(Y=1) = \boxed{0.9.}$$

3.

$$E[Y] = 115\cdot P(Y=115) + 20\cdot P(Y=20) = 115\cdot 0.9 + 20\cdot 0.1 = \boxed{105.5.}$$

Decision Trees

(a) **Train a Decision Tree**

$$\begin{aligned} H(Y|C) = &-\big(P(Y=No, C=1^{st})\log(P(C=1^{st}|Y=No)) \\ &+ P(Y=Yes, C=1^{st})\log(P(C=1^{st}|Y=Yes)) \\ &+ P(Y=No, C=Lower)\log(P(C=Lower|Y=No)) \\ &+ P(Y=Yes, C=Lower)\log(P(C=Lower|Y=Yes))\big) \\ &= -\left(\frac{122}{2201}\log\left(\frac{122}{1490}\right) + \frac{203}{2201}\log\left(\frac{203}{711}\right) + \frac{1368}{2201}\log\left(\frac{1368}{1490}\right) + \frac{508}{2201}\log\left(\frac{508}{711}\right)\right) \\ &= 0.555. \end{aligned}$$

$$
\begin{aligned}
H(Y|G) = &- (P(Y = No, G = Male) \log(P(G = Male|Y = No)) \\
&+ P(Y = Yes, G = Male) \log(P(G = Male|Y = Yes)) \\
&+ P(Y = No, G = Female) \log(P(G = Female|Y = No)) \\
&+P(Y = Yes, G = Female) \log(P(G = Female|Y = Yes))) \\
&= - \left(\frac{1364}{2201} \log \left(\frac{1364}{1490} \right) + \frac{367}{2201} \log \left(\frac{367}{711} \right) + \frac{126}{2201} \log \left(\frac{126}{1490} \right) + \frac{344}{2201} \log \left(\frac{344}{711} \right) \right) \\
&= 0.606.
\end{aligned}
$$

$$
\begin{aligned}
H(Y|A) = &- (P(Y = No, A = Child) \log(P(A = Child|Y = No)) \\
&+ P(Y = Yes, A = Child) \log(P(A = Child|Y = Yes)) \\
&+ P(Y = No, A = Adult) \log(P(A = Adult|Y = No)) \\
&+P(Y = Yes, A = Adult) \log(P(A = Adult|Y = Yes))) \\
&= - \left(\frac{52}{2201} \log \left(\frac{52}{1490} \right) + \frac{57}{2201} \log \left(\frac{57}{711} \right) + \frac{1438}{2201} \log \left(\frac{1438}{1490} \right) + \frac{654}{2201} \log \left(\frac{654}{711} \right) \right) \\
&= 0.278.
\end{aligned}
$$

Since it has the greatest conditional entropy, G is the best feature to place at the root of the decision tree. The decision stump predicts $Y = No$ if $G = Male$ and $Y = Yes$ if $F = Female$.

(b) **Evaluation**

1. The decision stump correctly predicts Y for 1708 of the 2201 samples. Thus, it has accuracy $\frac{1708}{2201} = \boxed{0.776.}$.
2. We add up the total number of correct predictions for each combination of C,G, and A to get 1713 correct predictions, for an accuracy of $\frac{1713}{2201} = \boxed{0.778.}$

(c) **Decision Trees and Equivalent Boolean Expressions**
In the following tree, a left edge indicates a feature being false (0), whereas a right child indicates a feature being true (1).

(d) **Model Complexity and Data Size**
Note that the boolean expression $x_1 \vee (\neg x_1 \wedge x_2 \wedge x_6)$ is equivalent to the boolean expression $x_1 \vee (x_2 \wedge x_6)$, to which we simplify it.

1. $P(Y = 1|x_1 \vee (x_2 \wedge x_6)) = P(Y = 1|f(x) = 1) = \boxed{\theta.}$

2. $P(Y = 1|\neg(x_1 \vee (x_2 \wedge x_6))) = P(Y = 1|f(x) = 0) = \boxed{1 - \theta.}$
3. No; $P(Y = 1|X_2 = 1) = \frac{3}{4} \neq \frac{5}{8} = P(Y = 1)$.
4. Yes; since Y is a function of $f(x)$ and whether or not $y = f(x)$, each of which is independent of X_4 so that $Y \perp X_4$, and thus $P(Y = 1|X_4 = 1) = P(Y = 1)$.
5. Since the classifier is identical to f, the probability that it correctly predicts Y is $P(Y = f(x)) = \theta$.
6. Since the classifier is identical to f, the probability that it correctly predicts Y is $P(Y = f(x)) = \theta$.
7. In order to perfectly learn f, the decision tree must take into account the values of x_1, x_2, and x_6, so that it must have height 3 (not counting the leaves).

Maximum Lilelihood and MAP Estimation

(a) **Maximum Likelihood Estimation**

1. $P(X_1 \dots X_n|\theta) = \theta^m(1-\theta)^{1-m}$, where $m = \sum_{i=1}^n X_i$.
2. See attached plot. The following code was used to generate the plot (up to formatting):

```
>> theta = 0:0.01:1;
>> p = (theta.^6).*((1 - theta).^3);
>> plot(theta,p);
```

3. The following code was used to determine the maxizing value of θ:

```
>> [~,i] = max(p);
>> theta(i)

ans =

    0.667
```

Thus, θ^{MLE} agrees with the closed-form maximum likelihood estimator:

$$\frac{\sum_{i=1}^n X_i}{n} = \frac{6}{9} \approx 0.667.$$

4. See attached plot. The following code was used to generate the plot (up to formatting):

```
>> theta = 0:0.01:1;
>> p = (theta.^2).*((1 - theta).^1);
>> plot(theta,p);
```

See attached plot. The following code was used to generate the plot (up to formatting):

```
>> theta = 0:0.01:1;
>> p = (theta.^40).*((1 - theta).^20);
>> plot(theta,p);
```

5. The likelihood curves become narrower as the data set becomes larger. The maximum likelihood ($P(X_1 \ldots X_n|\theta^{MLE})$) becomes smaller as the data set becomes larger. The maximum likelihood estimate (θ^{MLE}) remains constant at $\frac{2}{3}$.

(b) **MAP Estimation**

1. See attached plot. The following code was used to generate the plot (up to formatting):

```
>> theta = 0:0.01:1;
>> p = 30.*(theta.^2).*((1 - theta).^2);
>> plot(theta,p);
```

2. See attached plot. The following code was used to generate the plot (up to formatting):

```
>> theta = 0:0.01:1;
>> p = 30.*(theta.^8).*((1 - theta).^5);
>> plot(theta,p);
```

$\theta^{MAP} = \frac{8}{13} \neq \theta^{MLE}$.

3. Yes; pick $Beta(\theta; 5, 3)$. Then, the posterior distribution will be identical to the likelihood distribution for 6 heads and 3 tails, since we "hallucinate" an additional $5 - 1 = 4$ heads and $3 - 1 = 2$ tails, for a total of $2 + 4 = 6$ heads and $1 + 2 = 3$ tails.