CS 475 Machine Learning: Homework 1
Learning Foundations
Due: Monday September 15, 2014, 11:59pm
50 Points Total Version 1.0

Haoyuan Ji (haoyuanji@gmail.com)

1 (3 points) We want to evaluate the performance of a fraud detection system. Suppose we have 1000 transactions where 10 are actually fraud. After analyzing all these transactions, the fraud detection system flagged 12 of these transactions as fraud, which contains 8 actually fraud ones. Give the accuracy (the percentage of correctly labeled examples), recall and precision values of the system for this experiment respectively.

ANSWER:

$$Accuracy = \frac{correctly\ labeled\ transactions}{total\ number\ of\ transactions} = \frac{986+8}{1000} = 99.4\% \tag{1}$$

$$Precision = \frac{correctly\ detected\ fraud\ transactions}{total\ number\ of\ detected\ fraud\ transactions} = \frac{8}{12} = 66.7\% \tag{2}$$

$$Recall = \frac{correctly\ detected\ fraud\ transactions}{total\ number\ of\ fraud\ transactions} = \frac{8}{10} = 80\% \tag{3}$$

2 (7 points) J.R.R. Tolkien is famous for his classic high fantasy works: Lord of the Rings, The Hobbit, and The Silmarillion, among others. Fans of his books are likely to have read more than one. After a survey of 100 readers, we found the following statistics:

1. 75 people had read Lord of the Rings, 50 had read The Hobbit, and 25 had read The Silmarillion
2. Of those that read Lord of the Rings, 40 had read The Hobbit, and 20 had read The Silmarillion

Suppose I want to know if a person has read Lord of the Rings without asking them directly. Instead, I can ask them if they have read either The Hobbit or The Silmarillion. I want to ask the single question that will give me the most information about whether they have read Lord of the Rings. Which question should I ask and why? Justify your answer in terms of information gain.

ANSWER:

Let sample space $\mathbf{Y}$ = {have read Lord of the Rings, have not read Lord of the Rings}
Let sample space $\mathbf{X}_1$ = {have read The Hobbit, have not read The Hobbit}
Let sample space $\mathbf{X}_2$ = {have read Silmarillion, have not read Silmarillion}

$$H(Y) = -\sum_{i=1}^{n} p(y_i)\log(p(y_i)) = -0.75 * \log(0.75) - 0.25 * \log(0.25) = 0.8113 \tag{4}$$

If we ask: **If you have read The Hobbit?**. Then

$$\begin{aligned} H(Y|X_1) &= \sum_{x \in X_1} p(x_{1i}) H(Y|X_1 = X_{1i})) && (5)\\ &= -\sum_{x \in X_1} p(x) \sum_{y \in Y} p(y|x) \log(p(y|x)) && (6)\\ &= -0.5 * [(0.8 * \log 0.8 + 0.2 * \log 0.2) + (0.7 * \log(0.7) + 0.3 * \log(0.3)] && (7)\\ &= 0.8016 && (8) \end{aligned}$$

So , the information gain for this question shoule be:

$$IG(Y|X_1) = H(Y) - H(Y|X_1) = 0.0097 \quad (9)$$

Similarly, if we ask: **If you have read The Silmarillion?**. Then

$$\begin{aligned} H(Y|X_2) &= \sum_{x \in X_2} p(x_{2i}) H(Y|X_2 = X_{2i})) && (10)\\ &= -\sum_{x \in X_2} p(x) \sum_{y \in Y} p(y|x) \log(p(y|x)) && (11)\\ &= -0.25 * (0.8 * \log 0.8 + 0.2 * \log 0.2) - 0.75 * (0.267 * \log(0.267) + 0.733 * \log(0.733)) && (12)\\ &= 0.8083 && (13) \end{aligned}$$

So , the information gain for this question shoule be:

$$IG(Y|X_2) = H(Y) - H(Y|X_2) = 0.003 \quad (14)$$

So, in terms of the information gain, we should ask **If you have read The Hobbit?**.

3 (5 points) Suppose that I have a data set with N unique features. I add some new features to the data, bringing the total number of unique features to M. How does the generalization ability of a model change as a result? Discuss in terms of the bias-variance tradeoff.

ANSWER:

For training set, increase the complexity of the model may reduce the error of the results. That will make your predictions fit better in your training set.

However, increase the complexity of your model will increase the bias of your results in test sets while decrease the variance. There should be a saddle point for the degree of complexity to acheive the least error. So we should consider it in terms of bias-variance tradeoff.

If N is less than saddle point, then increase the number of features to M may achieve a better output(less error) because of lower variance. But if M is too large, then we will see the error of your model becomes worse because of higher bias.