CS578/STAT590: Introduction to Machine Learning **Fall 2014**

Problem Set 1

Gen Nishida *Handed In: September 6, 2014*

1 Review Questions

1. Assume the probability of getting *head* when tossing a coin is λ.

   - What is the probability of getting the first head at the (k+1)-th toss?

     $$(1-\lambda)^k\lambda$$

   - What is the expected number of tosses needed to get the first head?
     Let k be the number of tosses needed to get the first head. Then the expected number of k is defined by

     $$\begin{aligned} E[k] &= \lambda \times 1 + (1-\lambda)\lambda \times 2 + (1-\lambda)^2\lambda \times 2 + \cdots & (1)\\ (1-\lambda)E[k] &= (1-\lambda)\lambda + (1-\lambda)^2\lambda \times 2 + (1-\lambda)^3\lambda \times 3 + \cdots & (2) \end{aligned}$$

     By subtracting (2) from (1), we get

     $$\begin{aligned} \lambda E[k] &= \lambda + (1-\lambda)\lambda + (1-\lambda)^2\lambda + (1-\lambda)^3\lambda + \cdots \\ &= \lambda\frac{1}{1-(1-\lambda)} \\ &= 1. \end{aligned}$$

     Thus, $E[k] = 1/\lambda$.

2. Let $f(x,y) = 3x^2 + y^2 - xy - 11x$

   - What is the partial derivative of f with respect to x ($\frac{\partial f}{\partial x}$)? Find $\frac{\partial f}{\partial y}$ as well.

     $$\frac{\partial f}{\partial x} = 6x - y - 11, \;\; \frac{\partial f}{\partial y} = 2y - x$$

   - Find a point (x,y) that minimizes f.

     $$\begin{cases} 6x - y - 11 = 0 \\ 2y - x = 0 \end{cases}$$

     By solving these equations, we get $(x,y) = (2,1)$.

3. - Assume that $\omega \in R^n$ and b is a scalar. A hyperplane in $\mathbb{R}^n$ is the set $\{x : x \in \mathbb{R}^n, \omega^T x + b = 0\}$. For $n = 2$ and $n = 3$, draw on paper an example of a hyperplane. The hyperplane has it normal vector ω, and it is away from the origin by $-b/\|\omega\|$. The example of a hyperplane for $n = 2$ and $n = 3$ is in Figure 1.

Figure 1: Figure 1

- Assume we have two parallel hyperplanes: $\{x : x \in \mathbb{R}^n, \omega^T x + b_1 = 0\}$ and $\{x : x \in \mathbb{R}^n, \omega^T x + b_2 = 0\}$. What is the distance between these two hyperplanes?

$$\left| \frac{-b_1}{\|\omega\|} - \frac{-b_2}{\|\omega\|} \right| = \frac{|b_1 - b_2|}{\|\omega\|}$$

2 Basic Concepts

1. Define in one sentence: (1) training set, (2) test set, (3) validation set.

   - training set

     Training set is a set of data used to optimize a hypothesis function.

   - test set

     Test set is a set of real-world data used to measure the accuracy of the hypothesis generated through training and validation phases.

   - validation set

     Validation set is a set of data used to estimate the performance of the hypothesis.

2. Can you use the validation set as a test set?

   No. Since validation set is used to estimate the accuracy of the hypothesis during the validation phase, the resulting hypothesis is optimized for the validation set, and it is meaningless to use the validation set as a test set in order to measure the actual performance for the real-world data.

3. Define in one sentence: overfitting

   A hypothesis is said to overfit the training data if it has smaller error on the training data but loses the generalization performance and has larger error on test data.

4. True or False (and why): A learned hypothesis f has a training error e_{tr} and a testing error e_{ts}, where $e_{tr} > e_{ts}$.

   - (1) can we say that f overfits to the training data?

     False. Since the hypothesis f is optimized for the training data while the test data is unknown during training phase, the training error e_{tr} is smaller than e_{ts} in general, even if f is well generalized. In this case, $e_{tr} > e_{ts}$, which indicates that f is generalized very well.

   - (2) Now, assume that $e_{tr} < e_{ts}$, does f overfit to the training data?

     False. Since the hypothesis f is optimized for the training data while the test data is unknown during training phase, the training error e_{tr} is smaller than e_{ts} in general, even if f is well generalized. Therefore, we cannot conclude that f overfits to the training data even if $e_{tr} < e_{ts}$, unless we find another hypothesis f' which has larger error on the training data but smaller error on the test data compared to f.

3 Decision Trees

1. The "Thrill and Romance" bookstore

    - What is the entropy of the target variable? (Buy)
    - What are the
    - What is the first attribute
    - Due to a computer error

2. Decision Tree Implementation