Machine Learning & Data Mining
Homework 0

Caltech CS/CNS/EE 155
January 6th, 2015

Overview

- This homework is intended as a review of prerequisite materials.
- This homework does not need to be turned in and will not be graded.
- Students should be able to complete this homework in at most 5 hours, ideally around 2 hours.
- Students who cannot do most of this homework are STRONGLY ADVISED to drop the course.

1 Basics

Give a short 1-2 sentence answer to the following questions:

Question A: What is a hypothesis class? Specifically, what is the hypothesis class of a linear model?

Question B: Given finite training data, what is the most common way to prevent overfitting?

Question C: What is the difference between supervised learning and unsupervised learning?

Question D: What is the difference between training data and test data?

Question E: What is the fundamental training data sampling assumption that permits generalization?

Question F: What is the benefit of squared loss over hinge loss, and vice versa?

2 Maximum Likelihood & Logistic Regression

Suppose you have training data of the form $S = \{(x_i, y_i)\}_{i=1}^N$, where each $y_i \in \{0, 1\}$ is a binary training label, and each $x_i \in \Re^D$ is a D-dimensional feature vector. We have a probabilistic model $P(y|x, w)$ that predicts the probability of y given an x, and is parameterized by w.

The maximum likelihood formula for finding the optimal w is

$$\text{argmax}_w \prod_{(x_i,y_i)\in S} P(y_i|x_i, w) \tag{1}$$

Question A: Rewrite (1) as minimizing the negative log-likelihood. Argue that optimizing the two objectives lead to identical solutions.

Suppose our probabilistic model is the logistic regression model:

$$P(y|x, w) = \begin{cases} \frac{e^{w^T x}}{1+e^{w^T x}} & \text{if } y = 1 \\ \frac{1}{1+e^{w^T x}} & \text{if } y = 0 \end{cases}. \tag{2}$$

Question B: Is it typically easier to solve the maximum likelihood problem using (1) or the negative log-likelihood formulation? Why?

Question C: Write down the gradient descent formula for your answer to Question 2B.

3 Cross Validation

Suppose we had N training data points $S = \{(x_i, y_i)\}_{i=1}^N$, and would like to perform K-fold cross validation, for some integer K.

Question A: How many training sessions are there?

Question B: For each training session, how large are the training and validation sets?

Question C: Are training sets allowed to overlap during different sessions? Are validation sets?

Question D: Why would we want to do cross validation?

Question E: What is the largest possible value of K?

4 Bias-Variance Tradeoff

Consider a training set $S = \{(x_i, y_i)\}_{i=1}^N$. In this question, we will think of S as a random variable. That is to say, every time we collect a training set S, we can think of it as being drawn randomly from a distrubution (the exact nature of the distribution is not important).

Let $h_S(x) := y$ denote a model trained over the training set S. For every input x, $h_S(x)$ will predict a y. Note that h_S depends on the specific training set provided. If one were to collect multiple different training sets, each resulting h_S would be different.

In this question, we will be using squared error, $\ell(a, b) = (a - b)^2$, to evaluate the quality of any model h_S. First consider a single test point (x, y) randomly chosen from all of our test points. For a randomly collected training set S, we can write the expected loss of our model h_S as

$$E_S\left[(h_S(x) - y)^2\right], \tag{3}$$

where the expectation is taken over the randomness of S. The expected value of (3) (over the randomness of the randomly selected test point (x, y)) is also known as the generalization error of h_S.

Define $\bar{h} \equiv E_S[h_S(x)]$ as the average prediction of h_S on x over the randomness of the collected training data. It turns out that (3) is equivalent to

$$E_S[(h_S(x) - y)^2] = E_S\left[\left(h_S(x) - \bar{h}\right)^2\right] + \left(\bar{h} - y\right)^2. \tag{4}$$

Question A: Derive the (4) starting from (3). You should only need to use linearity of expectation and basic algebra to do this.

The first term in (4), $E_S\left[\left(h_S(x) - \bar{h}\right)^2\right]$, is known as the variance term. The second term in (4), $\left(\bar{h} - y\right)^2$, is known as the bias term.

Question B: Why are these two terms called the bias and variance terms?

Question C: Given a fixed-size training set, explain the bias-variance trade-off in terms of the two quantities in (4) and the choice of model class h_S. E.g., if the complexity of the model class increases, what are the implications of the two terms in (4)?

5 Bayesian Inference

This question pertains Bayesian posterior inference. Assume that we have three models in our model class, h_1, h_2, and h_3. For any given input x, each model h outputs a random binary label $h(x) \in \{-1, +1\}$. Figure 1 below shows the probability that each model h will output +1 for each of three inputs x_1, x_2, and x_3. Note

	x_1	x_2	x_3
h_1	1/4	1/4	3/4
h_2	1/2	1/4	1/4
h_3	0	1/2	1

Figure 1: Probablity of each model outputting $+1$ for each input, i.e., $P(h(x) = +1)$.

that the probability that each model h outputs -1 is just $P(h(x) = -1) = 1 - P(h(x) = +1)$. Assume that each model generates outputs completely independently for each input.

We receive observational data in the form of (x, y) where $x \in \{x_1, x_2, x_3\}$ and $y \in -1, +1$. This observational data is generated by one of the three models, and our goal is to infer the probability distribution of which of the three models is likely to be the model that's generating the observational data.

The prior probabilities of h_1, h_2, and h_3 are shown in Figure 2 below. That is to say, a priori, we believe

	Prior
h_1	1/2
h_2	1/4
h_3	1/4

Figure 2: Prior probabilities of each model.

that the true model is $1/2$ chance h_1, $1/4$ chance h_2, and $1/4$ chance h_3. Our goal is to update these beliefs as we receive observational data.

Question A: What is the updated probability distribution over h_1, h_2, and h_3 after observing $(x_1, -1)$?

Question B: What is the updated probability distribution over h_1, h_2, and h_3 after observing $(x_1, -1)$ and $(x_2, +1)$?

Question C: What is the updated probability distribution over h_1, h_2, and h_3 after observing $(x_1, -1)$, $(x_2, +1)$, and $(x_3, -1)$?