COMP 652 - Homework 1

Assigned: Sep 10, 2008
Due: Sep 17, 2008
Late: Sep 22, 2008

1. Explicit solution to univariate linear regression. (10 pts.)

Recall from class that the optimal weights for a least squares linear regression problem are $\mathbf{w} = (X^T X)^{-1} X^T Y$, where X is the input data matrix augmented with a column of ones, and Y is a column vector containing the output, or target, values:

$$X = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,n} & 1 \\ x_{2,1} & x_{2,2} & \dots & x_{2,n} & 1 \\ x_{3,1} & x_{3,2} & \dots & x_{3,n} & 1 \\ \vdots & \vdots & & \vdots & \vdots \\ x_{m,1} & x_{m,2} & \dots & x_{m,n} & 1 \end{bmatrix} \qquad Y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_m \end{bmatrix}$$

I said the formula above for $\mathbf{w}$ was *almost* an analytical solution to the problem, in the sense that it leaves us with a matrix inverse to perform. In general, there is no good formula for matrix inverses. But there is for the special case of a two-by-two matrix[1]. In univariate linear regression, the input vectors are of length one—that is, we have a single input variable, and we're looking at hypotheses of the form $h = w_0 + w_1 x$. In this case, $X^T X$ is a two-by-two matrix.

A. (7 pts.) Derive explicit formulae for the optimal weights w_0 and w_1 for a univariate linear regression problem, in terms of the x_i and the y_i. Show both your final formulae and how you arrived at those formulae.

B. (3 pts.) Use your formula on the data set below, which corresponds to the artificial univariate linear regression problem studied at the end of Lecture 1 and the start of Lecture 2, and check whether your answers for the optimal weights match those given in the lecture[2].

x	0.86	0.09	-0.85	0.87	-0.44	-0.43	-1.10	0.40	-0.96	0.17
y	2.49	0.83	-0.25	3.10	0.87	0.02	-0.12	1.81	-0.83	0.43

[1] If you do not remember this formula, look it up! Or, rederive it yourself by assuming a matrix $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ is given, and solving the system $\begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} \alpha & \beta \\ \gamma & \delta \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ for the unknowns α, β, γ and δ.

[2] Actually, the lecture reported $w_0 = 1.05$ and $w_1 = 1.60$. However, my recalculation just now resulted in $w_0 = 1.0588$ and $w_1 = 1.6102$.

2. Fitting polynomials and using cross-validation to avoid overfitting. (10 pts.)

Consider the data in file "COMP652_HW1_Q2Data.txt", which contains a number of (x, y) data instances.

A. (5 pts.) Implement polynomial regression for this data set. Show a plot of the data along with the optimal order-1 (i.e., linear), order-2, order-3 and order-4 fits to the data. Also, turn in your code.

B. (5 pts.) Implement leave-one-out cross validation, and use it to determine the order d that gives the fit with best estimated generalization error. State the best d, report the estimated generalization error for all d's tested, and turn in your code.

3. Error criterion for exponential noise. (10 pts.)

At the end of Lecture 2, we showed that one justification for minimizing the sum-squared-error (SSQ) criterion in a regression problem is that the hypothesis with least SSQ is also the one under which the data has maximum likelihood – if we assume that the target values are generated from the hypothesis, but perturbed by additive Gaussian noise. That is, if we assume that $y_i = h(\mathbf{x_i}) + e_i$, where the e_i are independent Gaussian random variables with standard deviation σ, then

$$\arg\max_{h \in \mathcal{H}} P(Y|X, h) = \arg\min_{h \in \mathcal{H}} \sum_{i=1}^{m} (y_i - h(\mathbf{x_i}))^2$$

Now, suppose that the noise variables were not Gaussian but rather exponentially distributed[3]. Recall that the exponential distribution has a single parameter, λ, and its density has the formula:

$$P_\lambda(t) = \begin{cases} \lambda e^{-\lambda t} & \text{if } t \geq 0 \\ 0 & \text{if } t < 0 \end{cases}$$

With this assumption about the e_i, the hypothesis that maximizes the likelihood of the data is no longer the one that minimizes the SSQ. Rather, it minimizes a different error function. In fact, it minimizes a different error criterion, subject to a certain constraint. Derive the error criterion and constraint for this case of exponentially-distributed noise. (Hint: the derivation showed in class for the Gaussian noise case mostly applies, but a few details are different.) Show your final result, as well as the derivation.

[3]Why this might be so, I don't know. But in general, different data sets may be subject to different kinds of noise, and exponential noise is convenient for this question.