

CS195-5: Introduction to Machine Learning
Fall 2006

Problem Set #1 : Apple

Out: September 14, 2006

Due: September 27, 2006, 11am EDT

Instructions

In order to identify problem sets to the automated submission system we will assign code names to each problem set. This one will be called “apple”.

How and what to submit? Please submit your solutions using the automated system in the following way.

1. Create a directory (say, `ps1`) in which you will work on your solutions. All files mentioned below are assumed to reside in that directory.
2. Typeset the written parts of your solution in \LaTeX . The final results should be a document in PDF (preferable) or in PostScript. please name this document `applesolution.pdf` or `applesolution.ps`, according to the format. **Important: please start the solution of every problem with `\newpage`.**
3. Create Matlab code (extension `.m`) and data (`.mat`) files needed.
4. Create a file named `README`, and include in it a short description of all Matlab files you are submitting.
5. **From the directory `ps1`**, run
`/courses/cs195-5/bin/cs195-5handin apple`

You can resubmit your work as many times as needed, and each subsequent version will override the previous one.

Late submissions: there will be a penalty of 25 points for any solution submitted past the deadline until 11am on Friday, September 29. No submissions will be accepted past then.

What is the required level of detail? When asked to derive something, please clearly state the assumptions, if any, and strive for balance: justify any non-obvious steps, but try to avoid superfluous explanations. When asked to plot something, please include the figure as well as the code used to plot it; if multiple entities appear on a plot, make sure that they are clearly distinguishable (by color or style of lines and markers). When asked to provide a brief explanation or description, try to make your answers concise, but do not omit anything you believe is important.

1 Regression

In this set of problems we will look at the regression problem and the maximum likelihood (ML) approach, with the goal to understand a bit better some of their properties.

We will start with simple linear regression, defined by

$$\hat{y} = f(\mathbf{x}; \mathbf{w}) = \mathbf{w}^T \mathbf{x} = \sum_{j=0}^d w_j x_j. \quad (1)$$

It is assumed in (1) that we have augmented the inputs $\mathbf{x} \in \mathbb{R}^d$ by adding 1 as the “zeroth” dimension, $x_0 \equiv 1$. This assumption is valid for the remainder of this assignment, unless stated otherwise.

Let $\hat{\mathbf{w}}$ be the linear regression parameters estimated using the least squares procedure from data $\mathbf{x}_1, \dots, \mathbf{x}_N$. We will denote by e_i the prediction error made by the resulting model on the i -th training example, i.e.

$$e_i = y_i - \hat{\mathbf{w}}^T \mathbf{x}_i.$$

It was stated in the lecture that an implication of the conditions imposed on $\hat{\mathbf{w}}$ is that the prediction errors e_i are *uncorrelated* with any linear function of the training data. To make this statement more precise, let us define the notion of correlation between two joint¹ scalar samples (set of values) $U = \{u_1, \dots, u_n\}$ and $V = \{v_1, \dots, v_n\}$. That correlation is defined as

$$\hat{\sigma}(u, v) = \sum_{i=1}^n (u_i - \bar{u})(v_i - \bar{v}), \quad (2)$$

¹We talk about “joint samples” in the sense that there is correspondence between u_i and v_i , in terms of order.

where

$$\bar{u} \triangleq \frac{1}{n} \sum_{i=1}^n u_i, \quad \bar{v} \triangleq \frac{1}{n} \sum_{i=1}^n v_i$$

are the *empirical sample means* (averages) of U and V , respectively. Intuitively, correlation measures how well one variable (say, u) is linearly predictable from the other (v).

Problem 1 [10 points]

Show that the prediction errors $y - f(\mathbf{x}; \hat{\mathbf{w}})$ are necessarily uncorrelated with any linear function of the training inputs. That is, show that for any $\mathbf{a} \in \mathbb{R}^{d+1}$,

$$\hat{\sigma}(e, \mathbf{a}^T \mathbf{x}) = 0.$$

End of problem 1

Next, we consider the effect of *scaling* the input in the regression problem. This becomes relevant, for instance, in polynomial regression with high order models, to prevent very large or very small values and the resulting potential for numerical instability (think about the case of $x = 0.01$ or $x = 1000$ for 10-order model...) It seems that a good solution could be to multiply each column of the design matrix \mathbf{X} by a suitable number so that the range of that column (i.e. of the corresponding dimension in the regression input space) be fixed, say, to $[-1, 1]$.

Suppose that the data are scaled by multiplying the j -th dimension of the input by a non-zero number c_j . We will denote a single normalized data point by $\tilde{\mathbf{x}} \triangleq [1, c_1 x_1, \dots, c_d x_d]^T$ and the resulting design matrix by $\tilde{\mathbf{X}}$.

Problem 2 [10 points]

Let $\hat{\mathbf{w}}$ be the ML estimate of the regression parameters from the unscaled \mathbf{X} , and let $\hat{\tilde{\mathbf{w}}}$ be the solution obtained from the scaled $\tilde{\mathbf{X}}$. Show that the scaling does not change optimality, in the sense that $\hat{\mathbf{w}}^T \mathbf{x} = \hat{\tilde{\mathbf{w}}}^T \tilde{\mathbf{x}}$.

End of problem 2

Advice: You may find it helpful to express the scaling as a linear operator, yielding a matrix-product expression, and to look at the matrix tutorial posted with Lecture 2.

In the lectures we have focused on estimation of \mathbf{w} . The other parameter of the model is the noise variance σ^2 . We can obtain a ML estimate of σ^2 in a similar way.

Problem 3 [5 points]

Derive the ML estimate of σ^2 under the Gaussian noise model.

End of problem 3

Advice: Just as we did for \mathbf{w} , compute the derivative with respect to σ^2 and find the value of σ^2 for which the derivative vanishes.

In the following problem we depart from the Gaussianity assumption, and look at a different noise distribution. Consider the statistical regression model

$$y = f(\mathbf{x}; \mathbf{w}) + \nu$$

in which random noise ν is distributed according to the following distribution:

$$p(\nu) = C \exp(-\nu^4), \quad (3)$$

where the constant C is set so as to ensure $\int_{-\infty}^{\infty} C \exp(-x^4) dx = 1$.

Problem 4 [10 points]

Derive the conditions on the maximum likelihood estimate $\hat{\mathbf{w}}_{ML}$ under this noise model. What loss function corresponds to this criterion? Using Matlab, plot this loss function L' for the interval $[-3, 3]$, and on the same axes plot the squared loss. Explain how do you expect the differences to affect the behavior of regression using one of these loss functions?

Now consider the following data set:

$$\begin{array}{c|ccc} x & -1 & 0 & 1 \\ \hline y & -1 & 1 & 1 \end{array}$$

We will find ML estimate of \mathbf{w} under the abovementioned model. However, instead of finding the solution in closed form, we will look for it numerically: For each combination w_0 and w_1 between -2 and 2, with a step of 0.01, calculate the empirical loss under the “non-Gaussian” L' , and find the settings of \mathbf{w} that correspond to the minimum. Plot the function obtained by this procedure as well as the function obtained with the linear least squares fit. Explain how what is seen on the plot relates to the differences in loss functions in the previous plot.

End of problem 4

Advice: The function testWs.m provided on the course website will be useful here; you just will need to define functions that calculate loss. To visualize the loss surface in Matlab, use function image(x,y,z).

In the remaining problem in this section we will explore temperature data obtained from the² Antarctic meteorological station Fort Gauss. The daily measurements collected at the station weekly in 2004 are provided in file `meteodata.mat`, in the matrix `data04`. The matrix includes one column per measurement. The first row contains the level of seismic activity in Richter units, the second row contains the reported temperature (in Kelvin).

Due to budget cuts, the station no longer has funding for thermometers. Instead, you were comisioned to develop a linear regression machine that will predict the temperatures from seismic activity, in an optimal way, based on the data from 2004. You will also have access to the data from 2005 in the matrix `data05` in the same file, but this is strictly to evaluate the performance of various models in the very end of the experiment. **Do not** use it to fit your models!

Problem 5 [10 points]

First, fit polynomial regression coefficients to the data using least squares for linear and quadratic regression (please use `degexpand.m` to generate the design matrix \mathbf{X}). Using 10-fold cross-validation, select the best model. Plot the resulting fit with the selected model, together with the data, and report the empirical loss (average sum of squares) and the log-likelihood of the data under the estimated Gaussian model. Based on this plot, briefly explain: do you have any misgivings about the use of Gaussian noise model?

End of problem 5

Further investigation has revealed that the lazy meteorologists measure the temperature inside the building, rather than doing it properly two miles from the station. This suggests an alternative model. In that model, the noise is drawn from exponential distribution

$$p(\nu) = \begin{cases} e^{-\nu} & \text{if } \nu > 0, \\ 0 & \text{otherwise.} \end{cases}$$

Problem 6 [10 points]

Perform 10-fold cross-validation evaluation of linear and quadratic regression under this new noise model, using numerical exhaustive search similar to that used in Problem 4 (again, please use `degexpand` to generate the design matrix). The range for all coefficients should be from -10 to 10 with step

²fictitious

0.5. Again, plot the resulting fit with the selected model, together with the data, and report the empirical loss (average sum of squares) and the log-likelihood of the data under the estimated exponential noise model. Which model (Gaussian or exponential) of the noise seems to be more adequate?

Now take the two models selected for Gaussian and exponential cases, and evaluate them on the test data from year 2005, and report which performs better in terms of likelihood and in terms of squared loss. Briefly state your conclusions from the whole experiment.

End of problem 6

2 Multivariate Gaussian distributions

In this set of problems we will be looking in more depth at multivariate Gaussian distributions. Recall that the probability density function (pdf) of a Gaussian distribution in \mathbb{R}^d is given by

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \mu)^T \Sigma^{-1}(\mathbf{x} - \mu)\right). \quad (4)$$

Problem 7 [10 points]

Show that a contour corresponding to a fixed value of pdf is an ellipse in the 2D x_1, x_2 space.

End of problem 7

Advice: You may find it easier to work in the log domain, and to write out explicitly the expression in (4) in terms of x_1 and x_2 .

We will now look at the maximum likelihood estimates for the parameters needed to describe a Gaussian distribution: its mean and covariance. Suppose we have a sample $\mathbf{x}_1, \dots, \mathbf{x}_n$ drawn from $\mathcal{N}(\mathbf{x}; \mu, \Sigma)$.

Problem 8 [10 points]

Show that the ML estimate of the mean vector μ is

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i$$

and the ML estimate of the covariance matrix Σ is

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^n (\mathbf{x}_i - \hat{\mu})(\mathbf{x}_i - \hat{\mu})^T.$$

End of problem 8

Advice: Start with deriving $\hat{\mu}$. You will find it helpful to express the Gaussian pdf in terms of $\mathbf{A} = \Sigma^{-1}$, and to use the following formulae for derivatives:

$$\frac{d}{d\mathbf{A}}(\mathbf{x} - \mu)^T \mathbf{A}(\mathbf{x} - \mu) = (\mathbf{x} - \mu)(\mathbf{x} - \mu)^T, \quad \frac{d}{d\mathbf{A}} \log |\mathbf{A}| = \mathbf{A}^{-1}.$$

3 Linear Discriminant Analysis

In this last set of problems we will consider an extension of linear discriminant analysis to multi-class case, and to nonlinear decision boundaries. You will need to download `apple_lda.mat`. The relevant variables in it are \mathbf{X} , a set of 2D points, and \mathbf{Y} , the class labels (one of three classes).

Let C be the number of classes; we will assume that the prior probability (estimated as the relative frequency) for each class is uniform, $1/C$. We will start with the assumption that the covariance matrices are identical for all the classes, i.e. $\forall k = 1, \dots, C, \Sigma_k = \Sigma$. The means μ_1, \dots, μ_C are of course distinct.

Problem 9 [10 points]

Show that the optimal decision rule is based on calculating a set of C linear discriminant functions

$$\delta_k(\mathbf{x}) = \mathbf{x}^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k \quad (5)$$

and selecting $C^* = \operatorname{argmax}_c \delta_c(\mathbf{x})$. Apply this method to the data, using ML estimate of means and covariances, and plot the resulting linear decision boundaries (or, alternatively, decision *regions*) along with the data. Report the classification error you obtain.

End of problem 9

Advice: A common trick to plot decision regions: create an indicator array, such that the value corresponds to class identity, over a relatively fine grid, and use the `image` command.

Problem 10 [10 points]

Now assume that the covariances are no longer required to be identical. Derive the (no longer linear) discriminant function for this case. Apply the resulting decision rule to the data, plot the decision boundaries/regions and report classification error. Briefly discuss your conclusions, including the reasons for difference in performance between the two models.

End of problem 10

Advice: The derivation is closely related to that in Problem 7.

Problem 11 [5 points]

An alternative approach to learning quadratic decision boundaries is to map the inputs into an extended feature space (five-dimensional in this case) with polynomial features (much like the extension of regression seen in Lecture 4). Implement that method, and compare the resulting decision boundaries to those obtained with quadratic discriminant functions in the previous problem. Are they identical/similar/qualitatively different? Explain the results of the comparison.

End of problem 11