ELEN 4903: Machine Learning

Columbia University, Spring 2016

Homework 1: Due February 5, 2016 by 11:59pm

Please read these instructions to ensure you receive full credit on your homework. Submit the written portion of your homework as a *single* PDF file through Courseworks (less than 5MB). In addition to your PDF write-up, submit all code written by you in their original extensions through Courseworks (e.g., .m, .r, .py, etc.). Any coding language is acceptable. Do not wrap your files in .rar, .zip, .tar and do not submit your write-up in .doc or other file type. Your grade will be based on the contents of *one* PDF file and the original source code. Additional files will be ignored. We will not run your code, so everything you are asked to show should be put in the PDF file. Show all work for full credit.

Late submission policy: Late homeworks will have 0.1% deducted from the final grade for each minute late. *Your homework submission time will be based on the time of your **last** submission to Courseworks. I will not revert to an earlier submission!* Therefore, do not re-submit after midnight on the due date unless you are confident the new submission is significantly better to overcompensate for the points lost. Submission time is non-negotiable and will be based on the time you submitted your last file to Courseworks. The number of points deducted will be rounded to the nearest integer.

Problem 1 (Maximum likelihood) – 30 points

Part 1. Imagine we have a sequence of N observations $(x_1, \ldots, x_N)$, where each $x_i \in \{0, 1\}$. We model this sequence as i.i.d. Bernoulli random variables, where $p(x_i = 1|\pi) = \pi$ and π is unknown.

(a) What is the joint likelihood of the data $(x_1, \ldots, x_N)$?

(b) Derive the maximum likelihood estimate $\hat{\pi}_{\text{ML}}$ for π.

(c) Explain why this maximum likelihood estimate makes intuitive sense.

Part 2. Now imagine another sequence of N observations $(x_1, \ldots, x_N)$, where each $x_i \in \{0, 1, 2, \ldots\}$. We model this sequence as i.i.d. Poisson random variables with unknown parameter λ. The following questions follow exactly from Part 1.

(a) What is the joint likelihood of the data $(x_1, \ldots, x_N)$?

(b) Derive the maximum likelihood estimate $\hat{\lambda}_{\text{ML}}$ for λ.

(c) Explain why this maximum likelihood estimate makes intuitive sense.

Problem 2 (Bayes rule) – 30 points

This problem builds on Part 2 of Problem 1. Again imagine we have a sequence of N non-negative integer-valued observations $(x_1, \ldots, x_N)$, which we model as i.i.d. Poisson random variables with unknown parameter λ. We place a gamma prior distribution on λ, written $\lambda \sim Gam(\lambda|a, b)$, where

$$Gam(\lambda|a, b) = \frac{b^a}{\Gamma(a)}\lambda^{a-1}e^{-b\lambda}.$$

(a) Use Bayes rule to derive the posterior distribution of λ and identify the name of this distribution.

(b) What is the mean and variance of λ under this posterior? Discuss how this relates to your solution to Part 2 of Problem 1.

Problem 3 (Linear regression) – 60 points

In this problem you will implement and analyze results using least squares linear regression. The goal is to predict the miles per gallon a car will get using six quantities about that car. The data can be found on the course website and on Courseworks.[1] The data set contains 392 instances of different car models, each containing a 6-dimensional feature vector (plus a dimension of 1's for the intercept) contained in X, and its average miles per gallon which we treat as the output contained in y.

Remember to include all original code with your homework .pdf submission.

Part 1. First, randomly split the data set into 20 testing examples and 372 training examples. Using the training data only, solve a linear regression model of the form $y \approx w_0 + \sum_{j=1}^{6} x_j w_j$ using least squares.

(a) Print the numbers you obtain for the vector $\hat{w}_{\text{ML}}$. Using the labels of each dimension contained in the readme file, explain what the sign of each value in $\hat{w}_{\text{ML}}$ says about the relationship of the inputs to the output.

(b) Use the least squares solution to predict the outputs for each of the 20 testing examples. Repeat this process of randomly splitting into training and testing sets 1000 times. Each time, calculate the mean absolute error of the resulting predictions, $\text{MAE} = \frac{1}{20}\sum_{i=1}^{20} |y_i^{\text{test}} - y_i^{\text{pred}}|$. What is the mean and standard deviation of the MAE for these 1000 tests?

Part 2. Using exactly the same training/testing setup as in Part 1, fit a pth order polynomial regression model using least squares for $p = 1, 2, 3, 4$. (Note that $p = 1$ is equivalent to Part 1.) For each value of p run 1000 experiments on randomly partitioned training/testing sets using 20 testing and 372 training examples. For each experiment calculate the root mean squared error,

$$\text{RMSE} = \sqrt{\frac{1}{20}\sum_{i=1}^{20}(y_i^{\text{test}} - y_i^{\text{pred}})^2}.$$

[1] See `https://archive.ics.uci.edu/ml/datasets/Auto+MPG` for more details on this dataset. Since I have done some preprocessing, you *must* use the data provided with this homework.

(a) In a table, print the mean and standard deviation of the RMSE as a function of p. Using these numbers argue for which value of p is the best.

(b) For each value of p, collect $y^{\text{test}} - y^{\text{pred}}$ for each test example. (Observe that this number can be negative, and there are 20×1000 in total.) Plot a histogram of these errors for each p.

(c) For each p, use maximum likelihood to fit a univariate Gaussian to the 20,000 errors from Part 2(b). Describe how you calculated the maximum likelihood values for the mean and variance (this is a univariate case of what we did in class, so no need to re-derive it). What is the log likelihood of these empirical errors using the maximum likelihood values for the mean and variance? Show this as a function of p and discuss how this agrees/disagrees with your conclusion in Part 2(a). What assumptions are best satisfied by the optimal value of p using this approach?