Homework 1: SOLUTIONS

Problem 1 (written) – 25 points

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. random variables from a Bernoulli distribution with unknown parameter $\pi \in [0, 1]$ and known parameter, where

$$p(x_i|\pi) = \pi^{x_i}(1-\pi)^{1-x_i}$$

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

$$\begin{aligned} p(x_1,\ldots,x_N|\pi) &= \prod_{i=1}^N p(x_i|\pi) && (1) \\ &= \prod_{i=1}^N \pi^{x_i}(1-\pi)^{1-x_i} && (2) \\ &= \pi^{\sum_{i=1}^N x_i}(1-\pi)^{\sum_{i=1}^N (1-x_i)} \quad \text{(not necessary to write this)} && (3) \end{aligned}$$

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

$$\hat{\pi}_{\text{ML}} = \arg\max_\pi p(x_1,\ldots,x_N|\pi) = \arg\max_\pi \prod_{i=1}^N p(x_i|\pi) = \arg\max_\pi \sum_{i=1}^N \ln p(x_i|\pi)$$

$$\frac{d}{d\pi}\sum_{i=1}^N \ln p(x_i|\pi) \;=\; \frac{1}{\pi}\sum_{i=1}^N x_i - \frac{1}{1-\pi}\sum_{i=1}^N (1-x_i) \;=\; 0$$

Solving gives

$$\hat{\pi}_{\text{ML}} = \frac{1}{N}\sum_{i=1}^N x_i$$

To help learn π, you use a prior distribution. You select the distribution $p(\pi) = \text{beta}(a, b)$.

(c) Derive the maximum a posteriori (MAP) estimate $\hat{\pi}_{\text{MAP}}$ for π?
SOLUTION:

$$\hat{\pi}_{\text{MAP}} = \arg\max_\pi p(\pi|x_1,\ldots,x_N) = \arg\max_\pi \frac{p(x_1,\ldots,x_N|\pi)p(\pi)}{p(x_1,\ldots,x_N)} = \arg\max_\pi \;\ln p(\pi) + \sum_{i=1}^N \ln p(x_i|\pi)$$

$$\frac{d}{d\pi}\ln p(\pi) + \sum_{i=1}^N \ln p(x_i|\pi) = \frac{a-1}{\pi} - \frac{b-1}{1-\pi} + \frac{1}{\pi}\sum_{i=1}^N x_i - \frac{1}{1-\pi}\sum_{i=1}^N (1-x_i) \;=\; 0$$

$$\hat{\pi}_{\text{MAP}} = \frac{a-1+\sum_{i=1}^N x_i}{a+b-2+N}$$

(d) Use Bayes rule to derive the posterior distribution of π and identify the name of this distribution.
SOLUTION:

$$p(\pi|x_1,\ldots,x_N) = \frac{p(x_1,\ldots,x_N|\pi)p(\pi)}{p(x_1,\ldots,x_N)} \propto p(\pi)\prod_{i=1}^{N} p(x_i|\pi)$$

$$p(\pi|x_1,\ldots,x_N) \propto \pi^{a-1}(1-\pi)^{b-1}\prod_{i=1}^{N}\pi^{x_i}(1-\pi)^{1-x_i} = \pi^{a-1+\sum_{i=1}^{N}x_i}(1-\pi)^{b-1+\sum_{i=1}^{N}(1-x_i)}$$

We recognize that this is proportional to a beta distribution,

$$p(\pi|x_1,\ldots,x_N) = beta(a',b'), \qquad a' = a + \sum_{i=1}^{N} x_i, \quad b' = b + \sum_{i=1}^{N}(1-x_i)$$

(e) What is the mean and variance of π under this posterior? Discuss how it relates to $\hat{\pi}_{\text{ML}}$ and $\hat{\pi}_{\text{MAP}}$.
SOLUTION:

Defining a' and b' as above, the posterior is $\pi \sim beta(a',b')$ and,

$$\mathbb{E}[\pi] = \frac{a'}{a'+b'}, \qquad Var(\pi) = \frac{a'b'}{(a'+b')^2(a'+b'+1)}$$

As the number of samples increases, this converges the to the ML and MAP solutions. The variance is decreasing (in general) as N increases, and it goes to zero as $N \to \infty$, meaning we become more and more confident in the ML/MAP solution the mean is converging to.

Problem 2 (coding) – 35 points

In this problem you will analyze data using the linear regression techniques we have discussed. The goal of the problem is to predict the miles per gallon a car will get using six quantities (features) about that car. The zip file containing the data can be found on Courseworks.[1] The data is broken into training and testing sets. Each row in both "X" files contain six features for a single car (plus a 1 in the 7th dimension) and the same row in the corresponding "y" file contains the miles per gallon for that car.

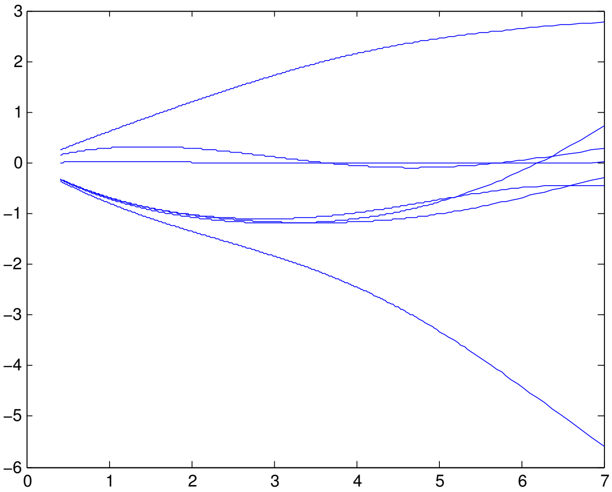
Remember to submit all original source code with your homework. Put everything you are asked to show below in the PDF file.

Part 1. Using the training data only, write code to solve the ridge regression problem

$$\mathcal{L} = \lambda \|w\|^2 + \textstyle\sum_{i=1}^{350} \|y_i - x_i^T w\|^2.$$

(a) For $\lambda = 0, 1, 2, 3, \ldots, 5000$, solve for w_{RR}. (Notice that when $\lambda = 0$, $w_{\text{RR}} = w_{\text{LS}}$.) In one figure, plot the 7 values in w_{RR} as a function of $df(\lambda)$. You will need to call a built in SVD function to do this (all details are in the slides). Be sure to label your 7 curves by their dimension in x.
SOLUTION:

Below is the correct figure using $df(\lambda)$



(a)

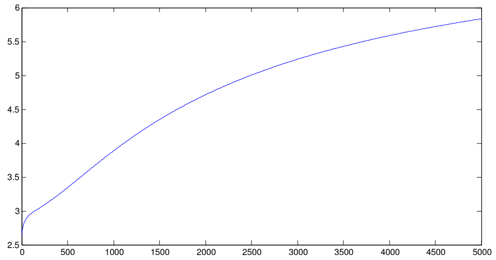
(b) The 4th dimension (car weight) and 6th dimension (car year) clearly stand out over the other dimensions. What information can we get from this?
SOLUTION:

The values of the 4th dimension are very negative. This indicates that as car weight increases, gas mileage decreases. The values of the 6th dimension is very positive. This indicates that newer cars tend to have much better gas mileage.

[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.

(c) For $\lambda = 0, \dots, 50$, predict all 42 test cases. Plot the root mean squared error (RMSE)[2] on the test set as a function of λ—*not* as a function of $df(\lambda)$. What does this figure tell you when choosing λ for this problem (and when choosing between ridge regression and least squares)?

SOLUTION:

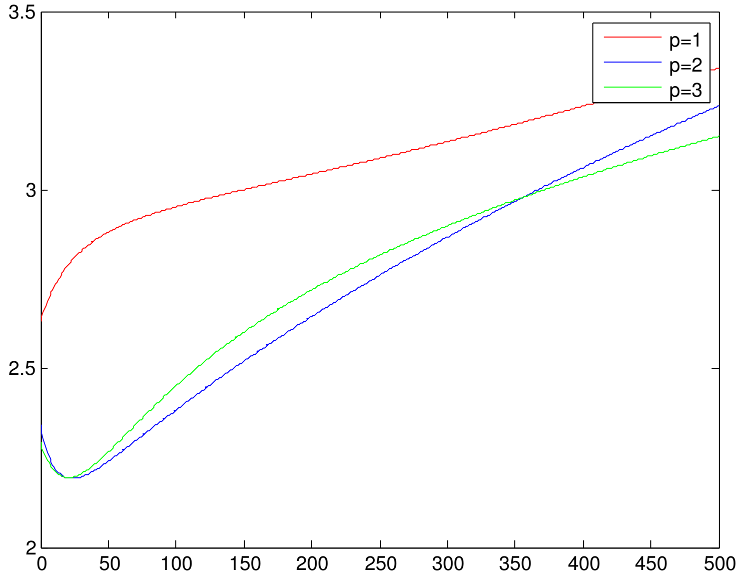


This plot indicates that $\lambda = 0$ performs the best. As a result we can say that least squares is better than ridge regression for this linear regression setup because $\lambda = 0$ is least squares and $\lambda > 0$ is linear regression.

Part 2. Modify your code to learn a pth-order polynomial regression model for $p = 1, 2, 3$. (You've already done $p = 1$ above.) For this implementation, do not include the cross terms for this problem, but instead use the method discussed in the slides.

(d) In one figure, plot the test RMSE as a function of $\lambda = 0, \dots, 500$ for $p = 1, 2, 3$. Based on this plot, which value of p should you choose and why? How does your assessment of the ideal value of λ change for this problem?

SOLUTION:



The values of p we should choose is $p = 2$ because the performance is best for $p = 2$ at the best value of λ (need to zoom in to see this). For $p = 2$, the best value is $\lambda \approx 23$.

[2] RMSE $= \sqrt{\frac{1}{42}\sum_{i=1}^{42}(y_i^{\text{test}} - y_i^{\text{pred}})^2}$.