COMS 4721: Machine Learning for Data Science

Columbia University, Spring 2023

Homework 1: Due Sunday, February 12, 2023 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, .c). Any coding language is acceptable, but do not submit notebooks. Also, do not wrap your files in .rar, .zip, .tar and do not submit your write-up in .doc or other file type. When resubmitting homeworks, please be sure to resubmit *all files*. 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 time!* 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 (written) – 25 points

Imagine you have a sequence of N observations $(x_1, \dots, x_N)$, where each $x_i \in \{0, 1, 2, \dots, \infty\}$. You model this sequence as i.i.d. from a Poisson distribution with unknown parameter $\lambda \in \mathbb{R}_+$, where

$$p(X|\lambda) = \frac{\lambda^X}{X!}\mathrm{e}^{-\lambda}$$

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

(b) Derive the maximum likelihood estimate λ_{ML} for λ.

To help learn λ, you use a prior distribution. You select the distribution $p(\lambda) = \text{gamma}(a, b)$.

(c) Derive the maximum a posteriori (MAP) estimate λ_{MAP} for λ?

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

(e) What is the mean and variance of λ under this posterior? Discuss how it relates to λ_{ML} and λ_{MAP}.

Problem 2 (written) – 20 points

(a) You have data (x_i, y_i) for $i = 1, \dots, n$, where $x \in \mathbb{R}^d$ and $y \in \mathbb{R}$. You model this as $y_i \overset{iid}{\sim} N(x_i^T w, \sigma^2)$. You use the data you have to approximate w with $w_{\text{RR}} = (\lambda I + X^T X)^{-1} X^T y$, where X and y are defined as in the lectures. Derive the results for $\mathbb{E}[w_{\text{RR}}]$ and $\mathbb{V}[w_{\text{RR}}]$ given in the slides.

(b) If w_{RR} is the ridge regression solution and w_{LS} is the least squares solution for the above problem, derive an equation for writing w_{RR} as a function of w_{LS} and the singular values and right singular vectors of feature matrix X. Recall that the singular value decomposition of $X = USV^T$.

Problem 3 (coding) – 30 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 $\lambda = 0, 1, 2, 3, \dots, 5000$, 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 as discussed in the slides. Be sure to label your 7 curves by their dimension in x.[2]

(b) Two dimensions clearly stand out over the others. Which ones are they and what information can we get from this?

(c) For $\lambda = 0, \dots, 50$, predict all 42 test cases. Plot the root mean squared error (RMSE)[3] 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)?

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 use the method discussed in the slides. Also, be sure to standardize each additional dimension of your data.

(d) In one figure, plot the test RMSE as a function of $\lambda = 0, \dots, 100$ 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?

[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.

[2] The dimensions correspond to: 1. cylinders, 2. displacement, 3. horsepower, 4. weight, 5. acceleration, 6. year made

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

Problem 1 (written) – 25 points

Imagine you have a sequence of N observations $(x_1, \ldots, x_N)$, where each $x_i \in \{0, 1, 2, \ldots, \infty\}$. You model this sequence as i.i.d. from a Poisson distribution with unknown parameter $\lambda \in \mathbb{R}_+$, where

$$p(X|\lambda) = \frac{\lambda^X}{X!} \mathrm{e}^{-\lambda}$$

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

(b) Derive the maximum likelihood estimate λ_{ML} for λ.

(a) $\mathcal{L} = \prod_{i=1}^{N} P(x_i|\lambda) = \prod_{i=1}^{N} \frac{\lambda^{x_i}}{x_i!} e^{-\lambda}$

(b) $\mathcal{L} = \prod_{i=1}^{N} P(x_i|\lambda)$

$$= \prod_{i=1}^{N} \frac{\lambda^{x_i}}{x_i!} e^{-\lambda}$$

$$\underset{\lambda}{\operatorname{argmax}}(\mathcal{L}) = \underset{\lambda}{\operatorname{argmax}}(\log(\mathcal{L}))$$

$$\ln(\mathcal{L}) = \sum_{i=1}^{N} x_i \ln(\lambda) - \sum_{i=1}^{N} \ln(x_i!) - \sum_{i=1}^{N} \lambda$$

$$= \ln(\lambda) \cdot \sum_{i=1}^{N} x_i - \cancel{\sum_{i=1}^{N} \ln(x_i!)} - N\lambda$$

$$\frac{d \log(\mathcal{L})}{d\lambda} = \frac{1}{\lambda} \sum_{i=1}^{N} x_i - N = 0$$

$$\frac{1}{\lambda} \sum_{i=1}^{N} x_i = N$$

$$\lambda_{ML} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

To help learn λ, you use a prior distribution. You select the distribution $p(\lambda) = \text{gamma}(a, b)$.

(c) Derive the maximum a posteriori (MAP) estimate λ_{MAP} for λ?

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

(e) What is the mean and variance of λ under this posterior? Discuss how it relates to λ_{ML} and λ_{MAP}.

(c). $\text{gamma}(a,b) = \dfrac{\lambda^{a-1} e^{-b\lambda} b^{a}}{(a-1)!}$

$$\lambda_{map} = \underset{\lambda}{\text{argmax}}\ P(X|\lambda) \cdot P(\lambda)$$

$$= \underset{\lambda}{\text{argmax}} \prod_{i=1}^{N} \frac{\lambda^{x_i}}{x_i!} e^{-\lambda} \cdot \frac{\lambda^{a-1} e^{-b\lambda} b^{a}}{(a-1)!}$$

$$= \ln(\lambda) \cdot \sum_{i=1}^{N} x_i - \sum_{i=1}^{N} \ln(x_i!) - N\lambda + \ln(\lambda^{a-1}) - b\lambda + \cancel{\ln(b^{a})} - \cancel{\ln((a-1)!)}$$

$$\frac{d}{d\lambda} = \frac{1}{\lambda} \sum_{i=1}^{N} x_i - N + \frac{a-1}{\lambda} - b = 0$$

$$\frac{1}{\lambda}\left(\sum_{i=1}^{N} x_i + (a-1)\right) = N + b$$

$$\lambda_{map} = \frac{\sum_{i=1}^{N} x_i + (a-1)}{N+b}$$

To help learn λ, you use a prior distribution. You select the distribution $p(\lambda) = \text{gamma}(a, b)$.

(c) Derive the maximum a posteriori (MAP) estimate λ_{MAP} for λ?

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

(e) What is the mean and variance of λ under this posterior? Discuss how it relates to λ_{ML} and λ_{MAP}.

d. $P(\lambda|X) \propto \lambda^{\sum_{i=1}^{N} x_i + a - 1} \cdot e^{-\lambda(N+b)}$

$$\approx \text{gamma}\left(\sum_{i=1}^{N} x_i + a,\ N+b\right)$$

e. $E[\lambda_{map}] = E\left[\frac{\sum_{i=1}^{N} x_i + (a-1)}{N+b}\right]$

$$= \frac{\sum_{i=1}^{N} x_i}{N+b} + \frac{a-1}{N+b}$$

$$\lambda_{map} = \frac{\sum_{i=1}^{N} x_i + (a-1)}{N+b}$$

$Var[\lambda_{map}] = E[\lambda_{map}^2] - E[\lambda_{map}]^2$

$$= 0$$

When $a=1$, $b=0$, $\lambda_{mL} = \lambda_{map}$

Problem 2 (written) – 20 points

(a) You have data (x_i, y_i) for $i = 1, \dots, n$, where $x \in \mathbb{R}^d$ and $y \in \mathbb{R}$. You model this as $y_i \overset{iid}{\sim} N(x_i^T w, \sigma^2)$. You use the data you have to approximate w with $w_{\text{RR}} = (\lambda I + X^T X)^{-1} X^T y$, where X and y are defined as in the lectures. Derive the results for $\mathbb{E}[w_{\text{RR}}]$ and $\mathbb{V}[w_{\text{RR}}]$ given in the slides.

(b) If w_{RR} is the ridge regression solution and w_{LS} is the least squares solution for the above problem, derive an equation for writing w_{RR} as a function of w_{LS} and the singular values and right singular vectors of feature matrix X. Recall that the singular value decomposition of $X = USV^T$.

(a)

$$\mathbb{E}[W_{RR}] = \mathbb{E}[(\lambda I + X^T X)^{-1} X^T y]$$

$$= (\lambda I + X^T X)^{-1} X^T X w$$

$$Var[W_{LS}] = \sigma^2 (X^T X)^{-1}$$

$$Var[W_{RR}] = Var[(\lambda (X^T X)^{-1} + I)^{-1} \cdot W_{LS}]$$

$$= (\lambda (X^T X^{-1}) + I)^{-1} \cdot Var[W_{LS}]$$

$$= (\lambda (X^T X^{-1}) + I)^{-1} \cdot \sigma^2 (X^T X)^{-1}$$

b. $X = USV^T \rightarrow (X^T X)^{-1} = VS^{-2}V^T$

$$W_{RR} = (\lambda (X^T X)^{-1} + I)^{-1} W_{LS}$$

$$= (\lambda V S^{-2} V^T + I)^{-1} W_{LS}$$

$$= V (\lambda S^{-2} + I)^{-1} V^T W_{LS}$$