Homework 1

Mingfeng Li
(ml4209)

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

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

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

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

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}.

a) $P(x_1, \ldots, x_N|\lambda) = \prod_{i=1}^{N} p(x_i|\lambda) = \prod_{i=1}^{N} \frac{\lambda^{x_i}}{x_i!} e^{-\lambda}$

b) $\ln \prod_{i=1}^{N} p(x_i|\lambda) = \sum_{i=1}^{N} \ln\left(\frac{\lambda^{x_i}}{x_i!} e^{-\lambda}\right) = \sum_{i=1}^{N} (x_i \ln\lambda - \ln x_i! - \lambda)$

$\nabla_\lambda \ln \prod_{i=1}^{N} p(x_i|\lambda) = 0$

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

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

$\bar{x} = \lambda_{ML}$

$\frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx}$

c) $\lambda_{MAP} = \arg\max_{\lambda} \ln p(\lambda|x_1, \ldots x_n)$

$= \arg\max_{x} \ln \frac{p(x_1, \ldots x_n|\lambda) p(\lambda)}{p(x_1, \ldots x_n)}$

$= \arg\max_{\lambda} \ln p(x_1, \ldots x_n|\lambda) + \ln p(\lambda)$

Let $L = \ln p(x_1, \ldots x_n|\lambda) + \ln p(\lambda)$

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

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

$=$

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

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

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

d) $$P(\lambda | x_1, \dots x_n) = \frac{P(x_1, \dots x_n | \lambda) P(\lambda)}{P(x_1, \dots x_n)}$$

$$\propto P(x_1, \dots, x_n | \lambda) P(\lambda)$$

$$\propto \lambda^{\sum_{i=1}^{n} x_i + a - 1} e^{-(n+b)\lambda}$$

$\Rightarrow P(\lambda | x_1, \dots x_n)$ is $\text{gamma}\left(\sum_{i=1}^{n} x_i + a, n+b\right)$

e) $$\hat{\lambda}_{post} = \frac{\sum_{i=1}^{n} x_i + a}{n+b} \qquad Var(\lambda_{post}) = \frac{\sum_{i=1}^{n} x_i + a}{(n+b)^2}$$

$$\hat{\lambda}_{post} = \frac{\sum_{i=1}^{n} x_i + a}{n+b} = \left(\frac{n}{n+b}\right) \frac{\sum_{i=1}^{n} x_i}{n} + \frac{a}{n+b}$$

$$= \frac{n}{n+b} \lambda_{ML} + \frac{a}{n+b}$$

$$\hat{\lambda}_{post} = \frac{\sum_{i=1}^{n} x_i + a}{n+b} = \frac{\sum_{i=1}^{n} x_i + a - 1}{n+b} + \frac{1}{n+b}$$

$$= \lambda_{MAP} + \frac{1}{n+b}$$

Problem 2 (written) – 15 points

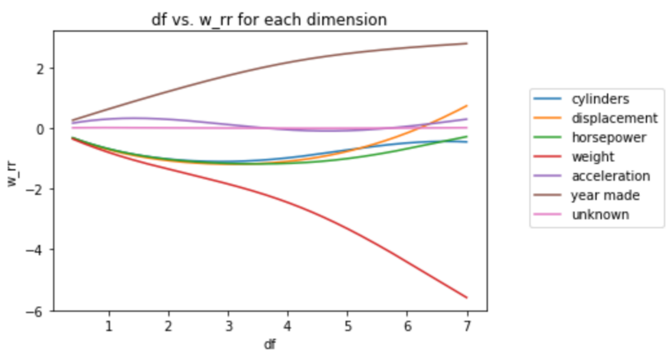
You have data (x_i, y_i) for $i = 1, \ldots, n$, where $x \in \mathbb{R}^d$ and $y \in \mathbb{R}$. You model this as $y_i \overset{iid}{\sim} N(x_i^T w, \sigma^2)$. You use the data you have to approximate w with $w_{\text{RR}} = (\lambda I + X^T X)^{-1} X^T y$, where X and y are defined as in the lectures.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

$$
\begin{aligned}
E[W_{rr}] &= E[(\lambda I + X^T X)^{-1} X^T y] \\
&= (\lambda I + X^T X)^{-1} X^T E[y] \\
&= (\lambda I + X^T X)^{-1} X^T X w
\end{aligned}
$$

$$
\begin{aligned}
Var[W_{rr}] &= E[(W_{rr} - E[W_{rr}])(W_{rr} - E[W_{rr}])^T] \\
&= E[W_{rr} W_{rr}^T] - E[W_{rr}]E[W_{rr}]^T \\
&= E[(\lambda I + X^T X)^{-1} X^T y y^T X (\lambda I + X^T X)^{-1}] - (\lambda I + X^T X)^{-1} X^T X w w^T X^T X (\quad)^{-1} \\
&= (\quad)^{-1} X^T (\sigma^2 I + \underline{X w w^T X^T}) X (\quad)^{-1} - (\quad)^{-1} X^T \underline{X w w^T X^T} X (\quad)^{-1} \\
&= (\quad)^{-1} X^T \sigma^2 I X [(\quad)^{-1}]^T \\
&= (\lambda I + X^T X)^{-1} X^T \sigma^2 I X [(\lambda I + X^T X)^{-1}]^T
\end{aligned}
$$

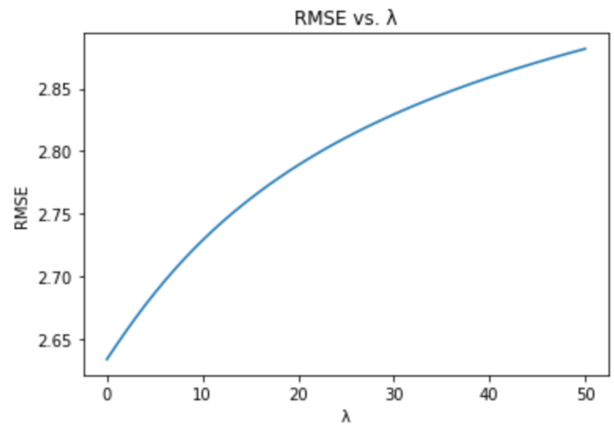
Problem 3.

Part I. a)



b) Year_Made and Weight clearly stands out from others. This suggests that as df ↗, $\lambda \to 0$, The w_rr getting closer to W_ls. Hence, Year_Made and Weight are two dimensions with significant W_ls value.

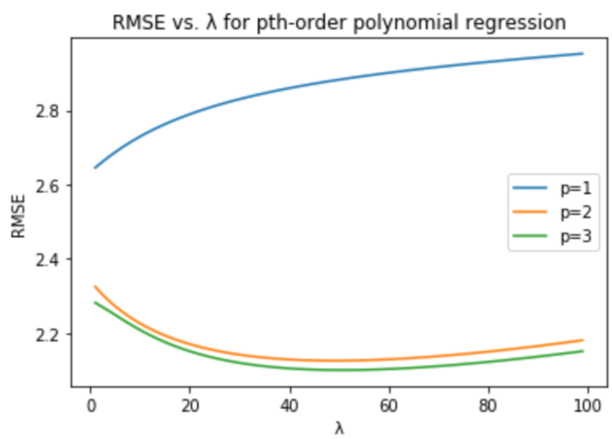
c)



The RMSE increases monotonicly as λ increases. This indicates that we would prefer a smaller λ value and we would prefer least squares to ridge regression.

Part II.

d)



According to the graph, I would choice value of $p = 2$, as 2nd-Order polynomial model have significant better (smaller RMSE) than 1st-Order and pretty close to 3-rd Order RMSE.

The RMSE decreases monotonically with λ increases from 0 to around 40. However, that large λ value would push w_rr closer to 0. A plot of w_rr with df for 2nd-Order Model would help here.