Problem 1

(a) $\mathcal{L}(\pi; x_1, \ldots, x_N) = \prod_{i=1}^{N} \pi^{x_i} (1-\pi)^{1-x_i}$ ✓

(b) $\ln \mathcal{L} = \sum_{i=1}^{N} [x_i \ln \pi + (1-x_i) \ln (1-\pi)]$

$$\frac{\partial \ln \mathcal{L}}{\partial \pi} = \sum_{i=1}^{N} \left(\frac{x_i}{\pi} - \frac{1-x_i}{1-\pi} \right)$$

At max, $\frac{\partial \ln \mathcal{L}}{\partial \pi} = 0 \Rightarrow \sum_{i=1}^{N} \frac{x_i}{\pi} = \sum_{i=1}^{N} \frac{1-x_i}{1-\pi}$

$$(1-\pi) \sum_{i=1}^{N} x_i = \pi \sum_{i=1}^{N} (1-x_i)$$

$$\sum_{i=1}^{N} x_i - \pi \sum_{i=1}^{N} x_i = N\pi - \pi \sum_{i=1}^{N} x_i$$

$$\sum_{i=1}^{N} x_i = N\pi$$

$$\pi = \frac{\sum_{i=1}^{N} x_i}{N} = \bar{x}$$

Since $\hat{\pi}_{ML} = \arg\max_{\pi} \mathcal{L} = \arg\max_{\pi} \ln \mathcal{L}$

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

(c) $\hat{\pi}_{MAP} = \arg\max_{\pi} P_r(\pi | X) = \arg\max_{\pi} \frac{P_r(X|\pi) p(\pi)}{P_r(X)}$

$$\Rightarrow \frac{P_r(X|\pi) p(\pi)}{P_r(X)} \propto P_r(X|\pi) p(\pi)$$

$$\ln \left(\mathcal{L}(\pi; X) p(\pi) \right) = \sum_{i=1}^{N} [x_i \ln \pi + (1-x_i) \ln (1-\pi)] + \ln \left[\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \pi^{a-1} (1-\pi)^{b-1} \right]$$

$$\frac{\partial}{\partial \pi} \ln \mathcal{L}(\pi; X) p(\pi) = \sum_{i=1}^{N} \left(\frac{x_i}{\pi} - \frac{1-x_i}{1-\pi} \right) + \frac{a-1}{\pi} - \frac{b-1}{1-\pi}$$

At max: $\sum_{i=1}^{N} \left(\frac{x_i}{\pi} - \frac{1-x_i}{1-\pi} \right) + \frac{a-1}{\pi} - \frac{b-1}{1-\pi} = 0$

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

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

$$= \pi \left(\sum_{i=1}^{N} x_i - \sum_{i=1}^{N} x_i + N + a + b - 2 \right)$$

$$\Rightarrow \hat{\pi}_{MAP} = \frac{\sum x_i + a - 1}{N + a + b - 2} \quad //$$

(d) $$Pr(\pi|X) = \frac{Pr(X|\pi)\,Pr(\pi)}{Pr(X)}$$

$$\propto Pr(X|\pi)\,Pr(\pi)$$

$$= \pi^{\Sigma x_i}(1-\pi)^{\Sigma(1-x_i)} \cdot \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\pi^{a-1}(1-\pi)^{b-1}$$

$$\propto \pi^{\Sigma x_i + a - 1}(1-\pi)^{\Sigma(1-x_i)+b-1}$$

This is the kernel of a $B(\Sigma x_i + a, N - \Sigma x_i + b)$ distribution.

After multiplying by the appropriate constant:

$$Pr(\pi|X) = \frac{\Gamma(N+a+b)}{\Gamma(\Sigma x_i + a)\Gamma(N-\Sigma x_i + b)}\pi^{\Sigma x_i + a - 1}(1-\pi)^{N-\Sigma x_i + b - 1}$$ //

i.e. beta $(\Sigma x_i + a, N - \Sigma x_i + b)$ //

(e) $$\text{Mean} = E(\pi) = \frac{\Sigma x_i + a}{\Sigma x_i + a + N - \Sigma x_i + b} = \frac{\Sigma x_i + a}{N + a + b}$$ //

$$\text{Var}(\pi) = \frac{(\Sigma x_i + a)(N - \Sigma x_i + b)}{(a+N+b)^2(a+N+b+1)}$$

The mean is the expected value of π under the posterior while the variance is a measure of the uncertainty in the parameter π.

This relates to $\hat{\pi}_{MAP}$ because they are both derived from the same beta $(\Sigma x_i + a, N - \Sigma x_i + b)$ distribution, except that $\hat{\pi}_{MAP}$ is the mode instead of the mean.

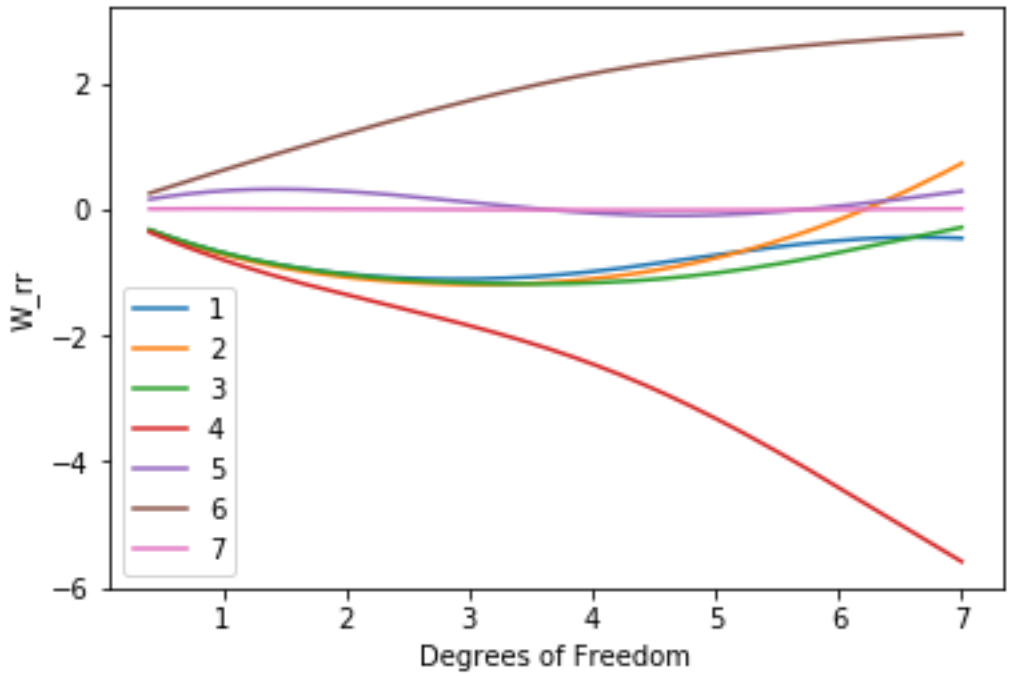
This relates to $\hat{\pi}_{ML}$ since MLE is a special case of MAP estimation that uses a uniform prior, i.e. a beta prior with $a, b = 1$.

Problem 2

Part 1

(a) For λ = 0,1,2,3, … ,5000, solve for w_{RR} . (Notice that when λ = 0, $w_{RR} = w_{LS}$.) In one figure, plot the 7 values in w_{RR} as a function of df(λ). 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.

In the figure below, the values for the ridge regression weights, w_{RR}, have been plotted against the degrees of freedom, $df(\lambda)$.

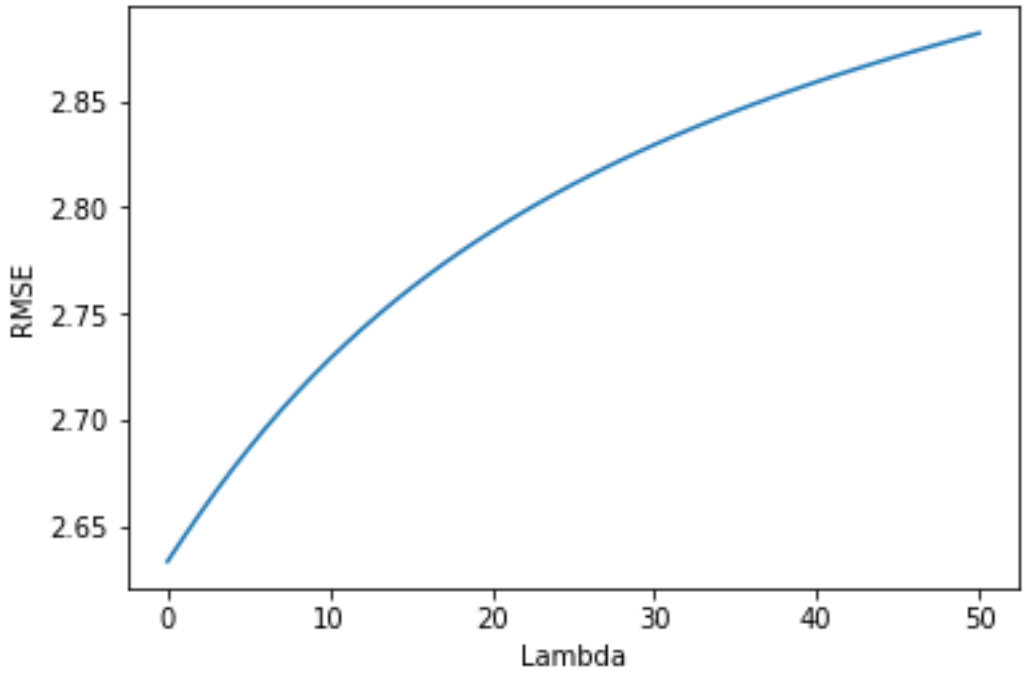


(b) The 4th dimension (car weight) and 6th dimension (car year) clearly stand out over the other dimensions. What information can we get from this?

From the graph above we can infer that dimension 4 (CAR WEIGHT) and dimension 6 (CAR YEAR) have the coefficients with the greatest magnitudes when λ is equal to 0. This means that the coefficients of these two features will be penalized the most by the ridge regression penalty term λ, since $df(\lambda)$ is an inverse function of λ.

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

In the figure below, the RMSE for the 42 predictions are plotted against λ.

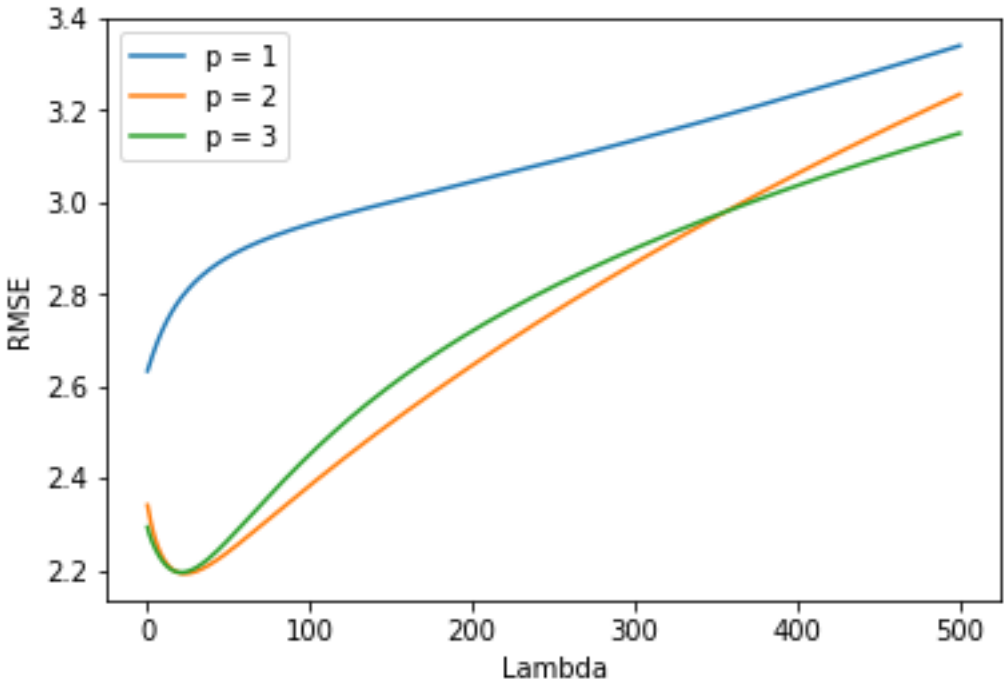


Since RMSE is at a minimum when $\lambda = 0$, we should choose 0 for our value of λ, which is equivalent to using a basic least squares model without the ridge regression penalty.

Part 2

(d) In one figure, plot the test RMSE as a function of $\lambda = 0, \ldots, 500$ 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?

In the figure below, the RMSE values of each of the three models are plotted against λ.



Based on this plot, we should choose p = 2 as this will result in the lowest RMSE value overall, but the p = 3 model seems to have a similar RMSE as well. The ideal value of λ changes from 0 (in the p = 1 model) to a nonzero value around 20-22.