ELEN E4903: MACHINE LEARNING HOMEWORK 1

Problem 1:

Given a sequence of N observations $(x_1, \cdots, x_N)$ with each $x_i \in \{0,1\}$ and $x_i \overset{iid}{\sim}$ Bernoulli(π) and $\pi \in [0,1]$,

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

(a) Joint likelihood of data: $p(x_1, \cdots, x_N|\pi) = \prod_{i=1}^{N} p(x_i|\pi)$

$$\therefore \boxed{p(x_1, \cdots, x_N|\pi) = \prod_{i=1}^{N} \pi^{x_i}(1-\pi)^{1-x_i} = \pi^{\sum_{i=1}^{N} x_i}(1-\pi)^{N-\sum_{i=1}^{N} x_i}}$$

(b) $\hat{\pi}_{ML} := \arg\max_{\pi} p(x_1, \cdots, x_N|\pi)$

$$= \arg\max_{\pi} \prod_{i=1}^{N} p(x_i|\pi)$$

Since taking the logarithm does not change the location of a maximum or minimum,

$$\hat{\pi}_{ML} = \arg\max_{\pi} \prod_{i=1}^{N} p(x_i|\pi)$$

$$= \arg\max_{\pi} \ln\left(\prod_{i=1}^{N} p(x_i|\pi)\right)$$

$$= \arg\max_{\pi} \sum_{i=1}^{N} \ln p(x_i|\pi)$$

Solving for $\hat{\pi}_{ML}$, $\nabla_\pi \sum_{i=1}^{N} \ln p(x_i|\pi) = \sum_{i=1}^{N} \nabla_\pi \ln p(x_i|\pi) = 0$

$$\sum_{i=1}^{N} \nabla_\pi \ln\left(\pi^{x_i}(1-\pi)^{1-x_i}\right) = 0$$

$$\sum_{i=1}^{N} \nabla_\pi \left[\ln \pi^{x_i} + \ln(1-\pi)^{1-x_i}\right] = 0$$

$$\sum_{i=1}^{N} \left[\frac{x_i}{\hat{\pi}_{ML}} + \frac{(1-x_i)}{1-\hat{\pi}_{ML}}(-1)\right] = 0$$

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

[Continued next page]

Continued

(b) $(1-\hat{\pi}_{ML})\sum_{i=1}^{N} x_i = \hat{\pi}_{ML}\sum_{i=1}^{N}(1-x_i)$

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

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

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

(c) Given prior $p(\pi) = \text{beta}(a,b)$, the MAP estimation seeks the most probable value π according to its posterior distribution

$\therefore \hat{\pi}_{MAP} = \arg\max_{\pi} \ln p(\pi \mid x_1, \ldots, x_N)$

$= \arg\max_{\pi} \ln \frac{p(x_1, \ldots, x_N \mid \pi) \cdot p(\pi)}{p(x_1, \ldots, x_N)}$ (Using Bayes' Rule)

$= \arg\max_{\pi} \ln p(x_1, \ldots, x_N \mid \pi) + \ln p(\pi) - \ln p(x_1, \ldots, x_N)$

Since the normalizing constant term $\ln p(x_1, \ldots, x_N)$ does not involve π, we can maximize the first two terms alone,

$\therefore \hat{\pi}_{MAP} = \arg\max_{\pi} \ln p(x_1, \ldots, x_N \mid \pi) + \ln p(\pi)$

$= \arg\max_{\pi} \sum_{i=1}^{N} \ln p(x_i \mid \pi) + \ln \text{beta}(\pi \mid a,b)$

Solving for $\hat{\pi}_{MAP}$, $\nabla_{\pi} \sum_{i=1}^{N} \ln\left(\pi^{x_i}(1-\pi)^{1-x_i}\right) + \nabla_{\pi} \ln\left\{\frac{\Gamma(a+b)}{\Gamma(a)\cdot\Gamma(b)} \cdot \pi^{a-1}(1-\pi)^{b-1}\right\}$

$\frac{\sum_{i=1}^{N} x_i}{\hat{\pi}_{MAP}} - \frac{\sum_{i=1}^{N}(1-x_i)}{1-\hat{\pi}_{MAP}} + \frac{a-1}{\hat{\pi}_{MAP}} - \frac{b-1}{1-\hat{\pi}_{MAP}} = 0$

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

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

$\hat{\pi}_{MAP}\left[N + b + a - 2\right] = \sum_{i=1}^{N} x_i + a - 1$

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

(d) According to Bayes' rule: $\underbrace{P(B|A)}_{\text{posterior}} = \underbrace{P(A|B)}_{\text{likelihood}} \cdot \underbrace{P(B)}_{\text{prior}} \div \underbrace{P(A)}_{\text{marginal}}$

In the context of this question,

$$\text{Posterior } p(\pi|x_1,\cdots,x_N) = \frac{p(x_1,\cdots,x_N|\pi)\cdot p(\pi)}{p(x_1,\cdots,x_N)}$$

$$= \frac{\left[\prod_{i=1}^{N} \pi^{x_i}(1-\pi)^{1-x_i}\right]\left[\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\pi^{a-1}(1-\pi)^{b-1}\right]}{p(x_1,\cdots,x_N)}$$

$$= \frac{\pi^{\sum_{i=1}^{N} x_i + a - 1}(1-\pi)^{\sum_{i=1}^{N}(1-x_i)+b-1}}{p(x_1,\cdots,x_N)} \cdot \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}$$

$$\propto \pi^{\sum_{i=1}^{N} x_i + a - 1}(1-\pi)^{\sum_{i=1}^{N}(1-x_i)+b-1} \quad \text{since the}$$

denominator $p(x_1,\cdots,x_N)$ and $\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}$ does not depend on π.

$\therefore$ We can recognize $\boxed{p(\pi|x_1,\cdots,x_N) = \text{beta}\left(\sum_{i=1}^{N} x_i + a, \sum_{i=1}^{N}(1-x_i)+b\right)}$

The posterior distribution of π is the beta distribution but with different parameters to the prior, that is, $\left(\sum_{i=1}^{N} x_i + a, \sum_{i=1}^{N}(1-x_i)+b\right)$ instead of (a,b).

(e) In (d), I found that π has a posterior distribution of beta $\left(\sum_{i=1}^{N} x_i + a, \sum_{i=1}^{N}(1-x_i)+b\right)$

∴ The mean of π under the posterior $= \dfrac{\sum_{i=1}^{N} x_i + a}{\sum_{i=1}^{N}(1-x_i)+b+\sum_{i=1}^{N} x_i + a}$

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

Variance of π under the posterior $= \dfrac{\left(\sum_{i=1}^{N} x_i + a\right)\left(\sum_{i=1}^{N}(1-x_i)+b\right)}{\left(\sum_{i=1}^{N} x_i + a + \sum_{i=1}^{N}(1-x_i)+b\right)^2\left(\sum_{i=1}^{N} x_i + a + \sum_{i=1}^{N}(1-x_i)+b+1\right)}$

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

The mean of π under the posterior is similar to $\hat{\pi}_{ML} = \frac{\sum_{i=1}^{N} x_i}{N}$ and $\hat{\pi}_{MAP} = \frac{\sum_{i=1}^{N} x_i + a - 1}{N+b+a-2}$ in that it seeks to obtain an estimate of π based on data that we have. It differs from $\hat{\pi}_{ML}$ in that it factors in a prior of $p(\pi) = \text{beta}(a,b)$ whereas ML only focuses on the likelihood and thus does not have the terms a and b. It differs from $\hat{\pi}_{MAP}$ in that it measures the mean while $\hat{\pi}_{MAP}$ is an estimate of the mode or most probable value of π under the posterior.

The variance of π under the posterior captures the uncertainty about π; just like what $\text{Var}(\hat{\pi}_{ML})$ and $\text{Var}(\hat{\pi}_{MAP})$ would have done. They are also similar in that with large N they asymptotically decay to 0.

ELEN E4903: MACHINE LEARNING HOMEWORK 1

Problem 2: Part 1

(a) $L = \lambda||w||^2 + \sum_{i=1}^{350} ||y_i - x_i^T w||^2$

$$w_{rr} = argmin_w \ \lambda||w||^2 + \sum_{i=1}^{350} ||y_i - x_i^T w||^2$$

$$= argmin_w \ (y - Xw)^T(y - Xw) + \lambda w^T w$$

$$\nabla_w L = -2X^T y + 2X^T X + 2\lambda w = 0$$

$$\therefore w_{rr} = (\lambda I + X^T X)^{-1} X^T y$$

For $\lambda = 0, 1, 2, 3, \ldots, 500$, a table of the first few values of w_{rr} is shown below where w1 corresponds to the estimate of w_{rr} for the first dimension of X and w7 corresponds to the estimate of w_{rr} for the seventh dimension of X:

Table 1: Estimates of w_{rr} for $\lambda = 0, 1, 2, 3, \ldots, 500$ (only the first 27 values are shown)

	Lambda	w1	w2	w3	w4	w5	w6	w7
1	0	-0.4562614	0.73016730	-0.2846187	-5.585589	0.2895777415	2.781398	1.015709e-02
2	1	-0.4457237	0.57776701	-0.3444969	-5.409686	0.2511063551	2.763335	8.127055e-03
3	2	-0.4413098	0.44574021	-0.3991785	-5.250289	0.2169052695	2.746405	6.362608e-03
4	3	-0.4414279	0.33021663	-0.4491997	-5.104980	0.1863707850	2.730449	4.815896e-03
5	4	-0.4449193	0.22825260	-0.4950435	-4.971818	0.1590096380	2.715338	3.450147e-03
6	5	-0.4509279	0.13756905	-0.5371403	-4.849223	0.1344139969	2.700969	2.236601e-03
7	6	-0.4588127	0.05637367	-0.5758722	-4.735891	0.1122432022	2.687253	1.152414e-03
8	7	-0.4680891	-0.01676267	-0.6115774	-4.630736	0.0922101518	2.674118	1.791921e-04
9	8	-0.4783868	-0.08299370	-0.6445549	-4.532843	0.0740709638	2.661501	-6.980582e-04
10	9	-0.4894211	-0.14326120	-0.6750693	-4.441433	0.0576170074	2.649351	-1.491690e-03
11	10	-0.5009716	-0.19834166	-0.7033548	-4.355841	0.0426686736	2.637621	-2.211984e-03
12	11	-0.5128667	-0.24888117	-0.7296194	-4.275493	0.0290704487	2.626274	-2.867567e-03
13	12	-0.5249726	-0.29542175	-0.7540476	-4.199890	0.0166869755	2.615274	-3.465743e-03
14	13	-0.5371844	-0.33842160	-0.7768037	-4.128598	0.0053998745	2.604593	-4.012736e-03
15	14	-0.5494202	-0.37827076	-0.7980343	-4.061236	-0.0048948442	2.594205	-4.513889e-03
16	15	-0.5616157	-0.41530339	-0.8178705	-3.997467	-0.0142889040	2.584086	-4.973816e-03
17	16	-0.5737206	-0.44980748	-0.8364296	-3.936993	-0.0228635165	2.574217	-5.396522e-03
18	17	-0.5856960	-0.48203260	-0.8538169	-3.879550	-0.0306908406	2.564579	-5.785506e-03
19	18	-0.5975117	-0.51219614	-0.8701271	-3.824902	-0.0378352043	2.555156	-6.143837e-03
20	19	-0.6091446	-0.54048836	-0.8854456	-3.772838	-0.0443541323	2.545934	-6.474220e-03
21	20	-0.6205774	-0.56707652	-0.8998494	-3.723166	-0.0502992150	2.536900	-6.779048e-03
22	21	-0.6317973	-0.59210828	-0.9134083	-3.675716	-0.0557168484	2.528042	-7.060448e-03
23	22	-0.6427953	-0.61571451	-0.9261856	-3.630332	-0.0606488642	2.519349	-7.320318e-03
24	23	-0.6535650	-0.63801161	-0.9382385	-3.586875	-0.0651330718	2.510813	-7.560356e-03
25	24	-0.6641029	-0.65910351	-0.9496193	-3.545216	-0.0692037238	2.502424	-7.782085e-03
26	25	-0.6744070	-0.67908325	-0.9603757	-3.505239	-0.0728919186	2.494175	-7.986880e-03
27	26	-0.6844770	-0.69803444	-0.9705509	-3.466837	-0.0762259503	2.486058	-8.175981e-03

Showing 1 to 28 of 5,001 entries

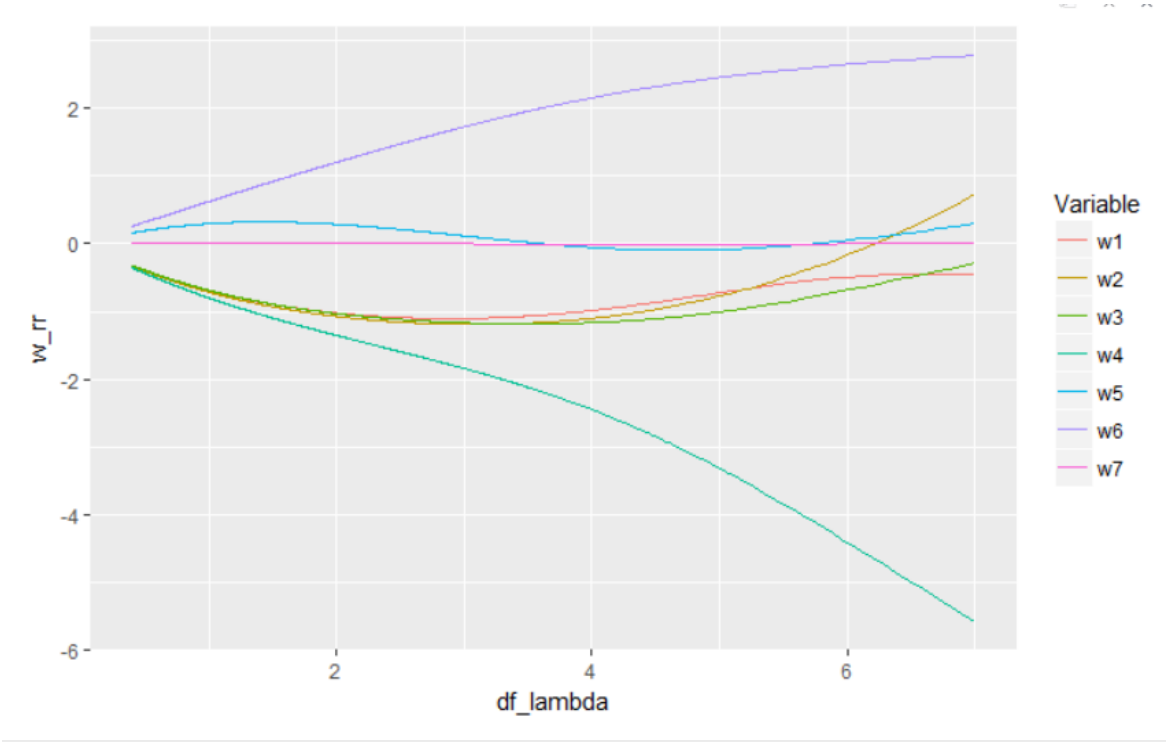
To obtain $df(\lambda)$, I obtained the singular value decomposition (SVD) of the matrix of features X so that $X = UDV^T \rightarrow (X^T X)^{-1} = VD^{-2}V^T$ and

$w_{rr} = (\lambda I + X^T X)^{-1} X^T y = V(\lambda D^{-2} + I)^{-1} V^T (X^T X)^{-1} X^T y = VMV^T (X^T X)^{-1} X^T y$ where M is a diagonal matrix with $M_{ii} = \frac{D_{ii}^2}{\lambda + D_{ii}^2}$ and

$$df(\lambda) = trace[(X(X^T X + \lambda I)^{-1} X^T] = \sum_{i=1}^{d} \frac{D_{ii}^2}{\lambda + D_{ii}^2} = trace(M)$$

Figure 1 below then plots the 7 values of w_{rr} as a function of $df(\lambda)$. Again, w1 corresponds to the estimate of w_{rr} for the first dimension of x and w7 corresponds to the estimate of w_{rr} for the seventh dimension of x:

Figure 1: Graph of w_{rr} against $df(\lambda)$



(b) The 4th dimension (car weight) and 6th dimension (car year) clearly stand out over the other dimensions as they have a larger value of w_{rr} for large values of $df(\lambda)$ (although they also decrease asymptotically to 0 as $df(\lambda)$ tends towards 0 as that corresponds to large values of λ and λ is the penalizing term for having a large value of w_{rr} such that when λ gets too large, it is optimal to set $w_{rr} = 0$). This implies that these 2 dimensions play a more significant role in explaining and predicting values of y (miles per gallon of a car) compared to the other dimensions in terms of magnitude.

The 4th dimension (car weight) takes on a significantly more negative value of w_{rr} than the other dimensions of X as shown in figure 2 above. This implies that for large $df(\lambda)$ or small values of λ (which gets us closer to least squares), we want to include car weight as an explanatory variable or predictor for y (miles per gallons of a car) since it exhibits a more significant negative relationship with y than the other variables. This makes sense as a heavier a car is the fewer miles per gallon we expect it to travel.

Conversely, the 6th dimension (car year) takes on a significantly more positive value of w_{rr} than the other dimensions of X as $df(\lambda)$ increases. This implies that for large $df(\lambda)$ or small values of

λ (which gets us closer to least squares), we want to include car year as an explanatory variable or predictor for y (miles per gallons of a car) since it exhibits a more significant positive relationship with y than the other variables. This makes sense as we expect newer cars to be more fuel efficient.

(c) For $\lambda = 0, \dots, 50$, using the w_{rr} obtained from the training data and applying it to the testing data, the predicted values of y, $\widehat{y_0}$, is given by $\widehat{y_0} = x_0^T w_{rr}$ where x_0 is the matrix of features in the testing data. A snapshot of the first 30 test cases of $\widehat{y_0}$ and for $\lambda = 0$ to 20 is shown in table 2 below:

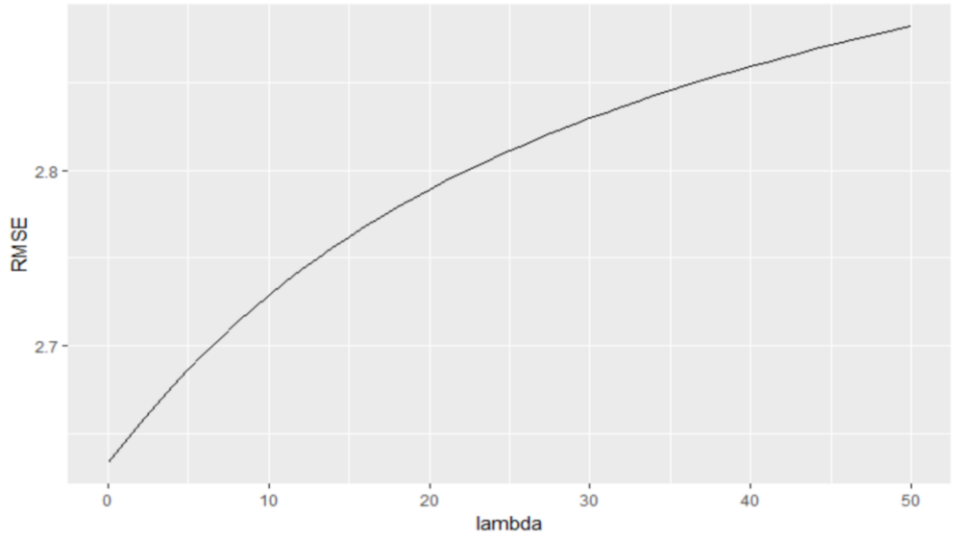
Table 2: Predicted values of y (only first 30 values (rows) and $\lambda = 0$ to 20 (columns) are shown)

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	-1.9088104	-1.8929834	-1.8778694	-1.8634512	-1.8496990	-1.8365768	-1.8240467	-1.8120703	-1.8006106	-1.78962248	-1.77910282	-1.76899094	-1.759268406	-1.74990898	-1.74088852	-1.73218482	-1.72377750	-1.71564780	-1.70777848	-1.70015369	-1.69275883
2	-9.6715327	-9.7448653	-9.8105416	-9.8696691	-9.9231380	-9.9716748	-10.0158801	-10.0562555	-10.0932247	-10.12714821	-10.15833519	-10.18705287	-10.213533485	-10.23798006	-10.26057102	-10.28146394	-10.30079854	-10.31869929	-10.33527744	-10.35063278	-10.36485516
3	7.2582487	7.2499274	7.2426671	7.2362211	7.2304035	7.2250725	7.2201183	7.2154553	7.2110161	7.20674719	7.20260599	7.19855831	7.194576700	7.19063902	7.18672741	7.18282746	7.17892756	7.17501834	7.17109232	7.16714354	7.16316729
4	-6.3349440	-6.2519698	-6.1743703	-6.1016402	-6.0333217	-5.9690037	-5.9083190	-5.8509403	-5.7965761	-5.74496646	-5.69587969	-5.64910859	-5.604467671	-5.56179056	-5.52092767	-5.48174422	-5.44411856	-5.40794055	-5.37311038	-5.33953731	-5.30713875
5	-0.5963691	-0.6954533	-0.7882580	-0.8754214	-0.9574971	-1.0349670	-1.1082525	-1.1777233	-1.2437054	-1.30648693	-1.36632382	-1.42344417	-1.478051932	-1.53033011	-1.58044341	-1.62854051	-1.67475601	-1.71921211	-1.76202001	-1.80328117	-1.84308838
6	2.2699266	2.2893408	2.3064484	2.3217032	2.3354452	2.3479341	2.3593714	2.3699156	2.3796935	2.38880727	2.39734063	2.40536251	2.412930398	2.42009260	2.42689003	2.43335762	2.43952542	2.44541939	2.45106217	2.45647356	2.46167098
7	-8.1193490	-8.1631708	-8.2044907	-8.2433558	-8.2798575	-8.3141104	-8.3462386	-8.3763689	-8.4046255	-8.43112749	-8.45598804	-8.47931320	-8.501202032	-8.52174662	-8.54103239	-8.55913849	-8.57613817	-8.59209922	-8.60708439	-8.62115177	-8.63435518
8	-1.8463352	-1.8809162	-1.9122060	-1.9406508	-1.9666079	-1.9903686	-2.0121742	-2.0322271	-2.0506993	-2.06773846	-2.08347221	-2.09801206	-2.111455983	-2.12389067	-2.13539324	-2.14603266	-2.15587087	-2.16496379	-2.17336203	-2.18111163	-2.18825455
9	-9.9992666	-10.0395014	-10.0753947	-10.1074913	-10.1362497	-10.1620586	-10.1852496	-10.2061071	-10.2248767	-10.24177131	-10.25697643	-10.27065446	-10.282948122	-10.29398331	-10.30387150	-10.31271174	-10.32059230	-10.32759210	-10.33378191	-10.33922534	-10.34397975
10	11.7976835	11.7363432	11.6785389	11.6238575	11.5719541	11.5225376	11.4753601	11.4302087	11.3868990	11.34527059	11.30518267	11.26651129	11.229146683	11.19299119	11.15795759	11.12396763	11.09095091	11.05884384	11.02758881	10.99713346	10.96743010
11	-7.1701457	-7.1381973	-7.1106519	-7.0866643	-7.0655893	-7.0469264	-7.0302810	-7.0153380	-7.0018429	-6.98958778	-6.97840143	-6.96814145	-6.958688579	-6.94994220	-6.94181691	-6.93423974	-6.92714807	-6.92048782	-6.91421212	-6.90828016	-6.90265625
12	8.4373873	8.4081616	8.3806197	8.3546185	8.3300266	8.3067242	8.2846030	8.2635655	8.2435237	8.22439840	8.20611828	8.18861896	8.171842229	8.15573543	8.14025078	8.12534487	8.11097818	8.09711466	8.08372135	8.07076810	8.05822722
13	-10.5363697	-10.5049182	-10.4770427	-10.4519978	-10.4292248	-10.4082977	-10.3888871	-10.3707345	-10.3536348	-10.33742355	-10.32196749	-10.30715781	-10.292904946	-10.27913459	-10.26578471	-10.25280323	-10.24014613	-10.22777605	-10.21566111	-10.20377397	-10.19209112
14	-2.3097498	-2.2981329	-2.2861530	-2.2741325	-2.2622766	-2.2507139	-2.2395220	-2.2287448	-2.2184035	-2.20850383	-2.19904182	-2.19000682	-2.181384082	-2.17315641	-2.16530528	-2.15781166	-2.15065652	-2.14382118	-2.13728758	-2.13103838	-2.12505708
15	-0.1157110	-0.1154489	-0.1148738	-0.1140058	-0.1128702	-0.1114943	-0.1099050	-0.1081278	-0.1061861	-0.10410143	-0.10189306	-0.09957826	-0.097172423	-0.09468918	-0.09214061	-0.08953738	-0.08688894	-0.08420358	-0.08148863	-0.07875056	-0.07599503
16	-15.4424678	-15.5092754	-15.5641821	-15.6094514	-15.6468208	-15.6776466	-15.7030021	-15.7237481	-15.7405823	-15.75407573	-15.76470020	-15.77284831	-15.778849229	-15.78298073	-15.78547858	-15.78654399	-15.78634953	-15.78504388	-15.78275567	-15.77959664	-15.77566418
17	-1.1290095	-1.0090920	-0.8995972	-0.7990744	-0.7063524	-0.6204718	-0.5406360	-0.4661769	-0.3965281	-0.33120528	-0.26979151	-0.21192497	-0.157289877	-0.10560902	-0.05663780	-0.01015934	0.03401948	0.07607132	0.11615113	0.15439847	0.19099942
18	1.8229043	1.8387688	1.8552274	1.8719289	1.8886364	1.9051900	1.9214825	1.9374432	1.9530272	1.96820711	1.98296867	1.99730636	2.011221051	2.02471800	2.03780553	2.05049400	2.06279508	2.07472126	2.08628544	2.09750063	2.10837980
19	10.0153152	9.9790748	9.9449708	9.9126814	9.8819542	9.8525873	9.8244172	9.7973089	9.7711504	9.74584715	9.72131878	9.69749642	9.674320513	9.65173922	9.62970712	9.60818416	9.58713485	9.56652757	9.54633403	9.52652880	9.50708893
20	-1.0136189	-1.0558706	-1.0909930	-1.1204553	-1.1453535	-1.1665203	-1.1845988	-1.2000930	-1.2134029	-1.22484999	-1.23469549	-1.24315366	-1.250401953	-1.25658858	-1.26183834	-1.26625707	-1.26993515	-1.27295026	-1.27536958	-1.27725150	-1.27864707
21	-13.7119942	-13.7189277	-13.7217967	-13.7215444	-13.7188663	-13.7142837	-13.7081935	-13.7009014	-13.6926457	-13.68361397	-13.67395516	-13.66378855	-13.653210329	-13.64229856	-13.63111699	-13.61971793	-13.60814450	-13.59643242	-13.58461136	-13.57270607	-13.56073725
22	11.0782135	11.0230011	10.9710678	10.9220217	10.8755361	10.8313360	10.7891873	10.7488898	10.7102703	10.67317826	10.63748228	10.60306671	10.569829401	10.53767972	10.50653689	10.47632863	10.44699005	10.41846268	10.39069365	10.36363504	10.33724329
23	-0.1091256	-0.1704436	-0.2250325	-0.2740147	-0.3182628	-0.3584672	-0.3951829	-0.4288619	-0.4598769	-0.48853868	-0.51510861	-0.53980865	-0.562828719	-0.58433245	-0.60446179	-0.62334062	-0.64107760	-0.65776860	-0.67349853	-0.68834296	-0.70236940
24	-9.8687842	-9.9496301	-10.0195274	-10.0805923	-10.1344206	-10.1822378	-10.2250005	-10.2634666	-10.2982442	-10.32982719	-10.35862106	-10.38496199	-10.409131252	-10.43136616	-10.45186846	-10.47081092	-10.48834246	-10.50459224	-10.51967292	-10.53368333	-10.54671054
25	-14.3533157	-14.5004049	-14.6278409	-14.7390375	-14.8366509	-14.9227822	-14.9991176	-15.0670272	-15.1276370	-15.18188130	-15.23054218	-15.27427975	-15.313655037	-15.34914802	-15.38117181	-15.41008391	-15.43619527	-15.45977765	-15.48106953	-15.50028110	-15.51759827
26	-8.2334457	-8.2306050	-8.2300663	-8.2311724	-8.2334491	-8.2365493	-8.2402162	-8.2442583	-8.2485314	-8.25292689	-8.25736223	-8.26177502	-8.266117977	-8.27035548	-8.27446091	-8.27841462	-8.28220238	-8.28581424	-8.28924354	-8.29248624	-8.29554033
27	3.6759567	3.6212023	3.5731804	3.5308051	3.4932243	3.4597581	3.4298558	3.4030649	3.3790099	3.35737553	3.33789477	3.32033955	3.304513647	3.29024705	3.27739158	3.26581732	3.25540982	3.24606774	3.23770098	3.23022911	3.22358004
28	1.4971469	1.4945323	1.4926465	1.4913989	1.4907094	1.4905076	1.4907322	1.4913294	1.4922524	1.49346051	1.49491785	1.49659325	1.498459304	1.50049199	1.50267016	1.50497519	1.50739065	1.50990201	1.51249640	1.51516243	1.51788997
29	-0.4723878	-0.4148651	-0.3624689	-0.3144213	-0.2701097	-0.2290425	-0.1908194	-0.1551102	-0.1216387	-0.09017205	-0.06051179	-0.03248735	-0.005951074	0.01922578	0.04315612	0.06593957	0.08766446	0.10840952	0.12824519	0.14723477	0.16543534
30	-1.6356514	-1.6340227	-1.6313771	-1.6279302	-1.6238469	-1.6192555	-1.6142565	-1.6089304	-1.6033416	-1.59754233	-1.59157545	-1.58547609	-1.579273363	-1.57299147	-1.56665062	-1.56026778	-1.55385721	-1.54743094	-1.54099914	-1.53457044	-1.52815214

Showing 1 to 32 of 42 entries

After obtaining the predicted values for all 42 test cases, the root mean squared error (RMSE) given by the equation $E[(y_0 - \widehat{y_0})^2] = E[(y_0 - x_0^T w_{rr})^2]$ is calculated and plotted as a function of λ as shown in Figure 2 below:

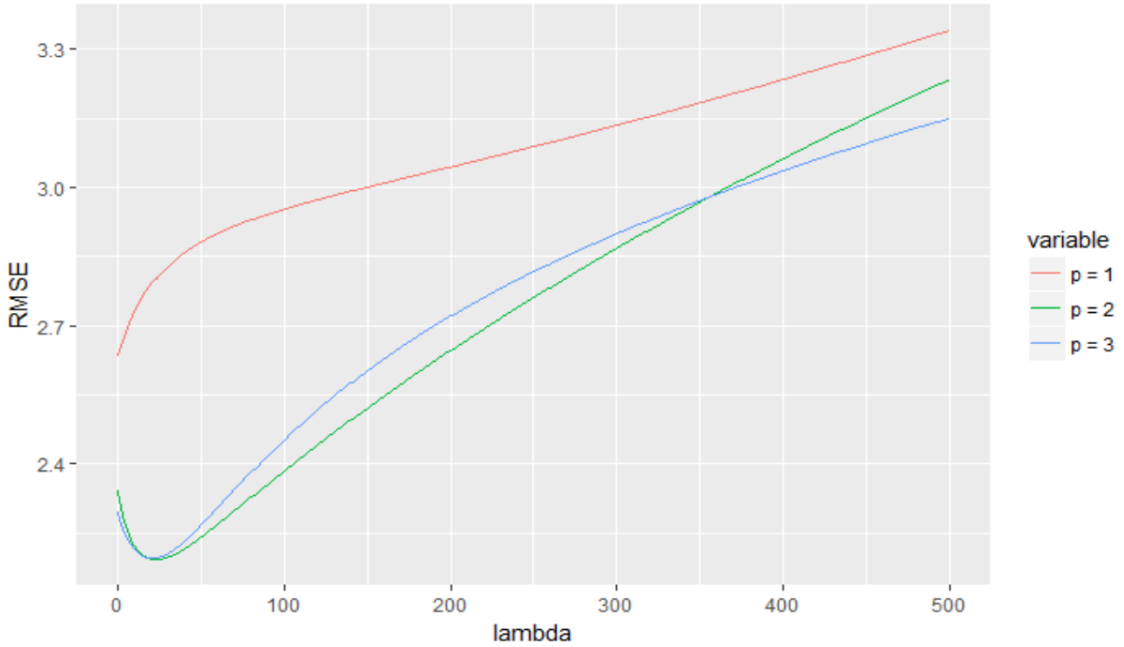
Figure 2: Graph of RMSE against λ



The fact that the RMSE is monotonically increasing as λ increases for $\lambda = 0, \dots, 50$ tells us that the optimal value of λ to choose is $\boldsymbol{\lambda = 0}$ since we want to minimize the RMSE as that minimizes the variance bias trade-off. This in turn implies that the **least squares model predicts the testing data better than a ridge regression** model in terms of **minimizing the RMSE**. This occurs because the **least squares model does not overfit the training data for $\boldsymbol{\lambda = 0, \dots, 50}$** so that there is no need for an additional penalizing factor λ that is present in a ridge regression. In fact, adding in λ to the objective function for a ridge regression model would result in an increase in bias that is more than the decrease in variance that the ridge regression brings about for the estimates and thus a poorer prediction accuracy for a ridge regression model compared to a least squares model.

(d) Figure 3 below shows the plot of RMSE as a function of For $\lambda = 0, \dots, 500$ and for p = 1, 2, 3.

Figure 3: Graph of RMSE against λ for p = 1, 2, 3



Finding the minimum point across all three different values of p yields the result that the value of p and λ that minimizes RMSE is **p = 2** and $\boldsymbol{\lambda = 23.}$ The reason why I would choose p = 2 is because the improvement in fit that it brings about in the testing data overweighs the increase in variance compared to the case when p = 1 thereby leading to an improvement in prediction accuracy in the testing data compared to p = 1. On the other hand, p = 3 performs more badly than p = 2 because it results in overfitting whereby the variance of the model increases so much that it outweighs the reduction in bias that p = 3 brings about. Hence **p = 2 gives the lowest RMSE as it optimizes the variance bias tradeoff.**

Now that we choose p = 2, the ideal value of $\boldsymbol{\lambda}$ is no longer 0 like in part c but is $\boldsymbol{\lambda = 23.}$ This is because with p = 2, we do not want the coefficients on the polynomials of X (w_{rr}) to be too large and want to impose a penalizing term in the ridge regression so as to maximize the likelihood over the hyperparameter $\boldsymbol{\lambda}$**.**