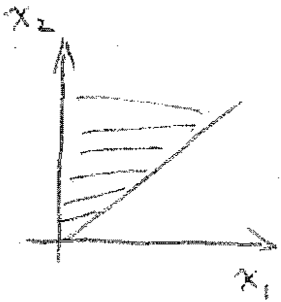
Homework 1 ML

Yuguang Li

1. Probability



1. $E(x) = \int_0^1 \int_0^1 \max(x_1, x_2)\ P(x_1, x_2)\ dx_1\, dx_2$

$$= \int_0^1 \int_0^{x_1} x_2\ dx_2\, dx_1 + \int_0^1 \int_0^{x_2} x_1\, dx_1\, dx_2$$

$$= \int_0^1 \frac{1}{2} x_2^2\, dx_2 + \int_0^1 \frac{1}{2} x_1^2\, dx_1$$

$$= \frac{1}{3}$$

2. $Var(x) = \int_0^1 \int_0^1 \left(\max(x_1, x_2) - \frac{1}{3}\right)^2 P(x_1, x_2)\ dx_1\, dx_2$

$$= \int_0^1 \int_0^1 \left((\max(x_1, x_2))^2 - \frac{2}{3}(\max(x_1, x_2)) + \frac{1}{9}\right) dx_1\, dx_2$$

$$= \int_0^1 \int_0^{x_1} x_2^2\, dx_2\, dx_1 + \int_0^1 \int_0^{x_2} x_1^2\, dx_1\, dx_2 - \frac{1}{9}$$

$$= \int_0^1 2 \cdot \frac{1}{3} x_1^3\, dx_1 - \frac{1}{9}$$

$$= \frac{1}{6} - \frac{1}{9} = \frac{1}{12}$$

3. $Cov(x, x_1) = \int_0^1 \int_0^1 (\max(x_1, x_2) - x_1)(\max(x_1, x_2) - x_2)\, dx_1 dx_2$

$$= \int_0^1 \int_0^1 (\max(x_1, x_2))^2 - (x_1 + x_2)\max(x_1, x_2) + x_1 x_2\ dx_1 dx_2$$

$$= \frac{1}{6} - \int_0^1 \int_0^1 (x_1 + x_2)\max(x_1, x_2) + x_1 x_2\ dx_1 dx_2$$

$$= \frac{1}{6} - \Big(\int_0^1 \int_0^{x_1} (x_1 + x_2) \cdot x_2 + x_1 x_2\ dx_2 dx_1 + \int_0^1 \int_0^{x_2} (x_1 + x_2)$$

$$\cdot x_1 + x_1 x_2\, dx_1 dx_2 \Big)$$

$$= \frac{1}{6} - \int_0^1 \frac{8}{3} x_1^3\, dx_1$$

$$= -\frac{1}{2}$$

2. 1. $P(x|\lambda) = \frac{\lambda^{x}}{x!} e^{-\lambda}$

$$P(G|\lambda) = \frac{\lambda^{\Sigma G_i}}{\prod G_i!} e^{-n\lambda}$$

$$\ln(P(G/\lambda)) = \ln\left(\frac{\lambda^{\Sigma G_i}}{G_i!} e^{-n\lambda}\right)$$

$$= \Sigma G_i \ln\lambda - \ln(\prod G_i!) - n\lambda$$

$$= \Sigma(G_i \ln\lambda - \ln G_i!) - n\lambda$$

2. $\frac{d\ln(P(G/\lambda))}{d\lambda} = \frac{\Sigma G_i}{\lambda} - n = 0 \qquad \lambda = \frac{n}{\Sigma G_i}$

3. $\frac{4+1+\cdots+8}{\lambda} - 8 = \frac{25}{\lambda} - 8$

$$\lambda_{MLE} = 8/25 = 0.32$$

3 1. ~~[illegible]~~ Complexity for X^TX

$O(n^2 d)$

Complexity for $(X^TX)^{-1}$

$O(n^2 d + d^3)$

2. $\hat{y}_i = x_i \hat{w} = x_i (X^TX)^{-1} X^T Y$

$= h_i Y$

3. $SSE_Z = (Y - Z_j)^T (Y - Z_j)$

$= \sum_{j=1}^{n} \sum_{i=1, i \neq j}^{n} (y_i - Z_j)^2 + \sum_{j=1}^{n} (y_i - Z_j)^2$

$= \sum_{j=1}^{n} \sum_{i=1, i \neq j}^{n} (y_i - y_i)^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i^{(-i)})^2$

$= \sum_{i=1}^{n} (y_i - \hat{y}_i^{(-i)})^2 = LOOCV$

4. $\hat{y}_i^{(-i)} = x_i \hat{w}^{(-i)} = x_i \cdot (X^TX)^{-1} X Z$

$= h_i \cdot Z$

5. $\hat{y}_i - \hat{y}_i^{(-i)} = \hat{y}_i - h_i Z$

$= \sum_i h_i y_i - h_{ii} \hat{y}_i^{(-i)} - \sum_{j \neq i} h_{ij} y_j$

$= h_{ii} y_i - h_{ii} \hat{y}_i^{(-i)}$

6. $O(3d + 3n)$

		ridge	LASSO	
4.	1. (a)	↘	↘	• Less Constraints on weights means more likely overfitting on training dataset.
	(b)	↗	↗	• Overfitting on traing dataset will make the model less accurate. So we will observe higher error on testing dataset
	(c)	↗	↗	• Less Constraints on weights will leave weights with high values.
	(d)	↗	↗	• Also more non-zeros elements of $\hat{w}$

• Because the regularization of two regression methods are similar. (L1 & L2 norm). So we will observe similar effects.

2. (a) Both ↗, because this ~~will~~ constrain the weights to be small. The model will be constrained to be simple. So the training error should ~~be~~ increase due to the simplicity of the model

(b) Both could be <u>either</u> increase or decrease. The over simplified model might be not good enough to describe the relation, which will result in the increase of testing error. On the other hand, a large λ might prevent overfitting. So the testing error might decrease

(c) Both decrease. A larger λ will control $\hat{w}$ to be small

(d) Both decrease same as (c)

5. 5.1 1. $L(f) = E_{X,Y}((Y - f(x))^2)$

$= E_{X,Y}((Y - E(Y|X) + E(Y|X) - f(x))^2)$

where $E(Y|X)$ is the conditional expectation of Y given X

$L(f) = E_{X,Y}(Y^2 - 2E(Y|X)Y + E(Y|X)^2) + E_{X,Y}(2E(Y|X)Y + 2E(Y|X)f(x)$

$- 2E(Y|X)^2 \cancel{A} - 2f(x)Y) + \cancel{\#} E_{X,Y}(E(Y|X)^2 - 2E(Y|X)f(x) + f(x)^2)$

$= E_{X,Y}((Y - E(Y|X))^2) + E_{X,Y}((E(Y|X) - f(x))^2)$

$+2E_{X,Y}((E(Y|X) - f(x))(Y - E(Y|X)))$

Because $E(Y|X) - f(x)$ and $Y - E(Y|X)$ are independent random variables

$E_{X,Y}((E(Y|X) - f(x))(Y - E(Y|X))) = 0$

So $L(f) = E_{X,Y}((Y - E(Y|X))^2) + E_X((E(Y|X) - f(x))^2)$

2. $L(f) = E_X(E(Y|X) - f(x))^2) + E_{X,Y}((Y - E(Y|X))^2)$

In order to minimize $L(f)$ we can have $L(f_{Bayes}) = E(Y|X$

while we can't do anything to the second term

5.2 $E_\tau L(\hat{f}) = E_\tau E_{XY}((Y - E(Y|X))^2) + E_\tau E_X((E(Y|X) - \hat{f}(x))^2)$

$= \cancel{E_\tau} E_{XY}((Y - E(Y|X))^2) + E_\tau E_X((E(Y|X)^2 + \bar{f}(x)^2 - 2E(Y|X)\bar{f}(x)$

$+ (\hat{f}(x)^2 - 2\bar{f}(x)\hat{f}(x) + \bar{f}(x)^2) + 2(\bar{f}(x)\hat{f}(x) + E(Y|X)\bar{f}(x)$

$- \bar{f}(x)^2 - 2E(Y|X)\hat{f}(x))$

$= E_{X,Y}((Y - E(Y|X))^2) + E_X((E(Y|X) - \bar{f}(x))^2)$

$+ E_\tau E_X((\hat{f}(x) - \bar{f}(x))^2) + 2E_\tau E_X((\bar{f}(x) - E(Y|X))(\hat{f}(x) - \bar{f}(x)))$

Because $\hat{f}(x) - \bar{f}(x)$ and $\bar{f}(x) - E(Y|x)$ are two independent random variable.

$$E_{\mathcal{Z}} E_x\left((\bar{f}(x) - E(Y|x))(\hat{f}(x) - \bar{f}(x)) \right) = 0 .$$

So $E_{\mathcal{Z}} L(\hat{f}) = E_{x,Y}\left((Y - E(Y|x))^2 \right) + E_x\left((E(Y|x) - \bar{f}(x))^2 \right)$
$+ E_{\mathcal{Z}} E_x\left((\hat{f}(x) - \bar{f}(x))^2 \right)$

2. $E_{x,Y}\left((Y - E(Y|x))^2 \right)$ noise
$E_x\left((E(Y|x) - \bar{f}(x))^2 \right)$ variance.
$E_{\mathcal{Z}} E_x\left((\hat{f}(x) - \bar{f}(x))^2 \right)$ bias

3. No f^* is the best estimator of $\bar{f}(x)$ it might be biased

4. With a large sample size N.
- $\bar{f}(x)$ is the theoretical Y value ~~&~~ at point X
- $E(Y|x)$ is the observed mean of Y given X
- $\hat{f}(x)$ is the estimated Y value using our model at point X. Our model is biased for most of the case (or is just optimized, but not the real one).

6 번 1. threshold 0.5

0/1 loss training : 0.0755

square loss training : 0.0505

2. 0/1 loss test : 0.0764

square loss test : 0.0515

3. Because the relation between features & labels might not be linear. Plus there are so many points lying around 0.5, which does not represent probability.

4. $\hat{y}_i = w^T x_i \leftarrow$ Bayes optimal predictor.

hard predictor $\hat{y}_i = \begin{cases} 0 & w^T x_i \le 0.5 \\ 1 & w^T x_i > 0.5 \end{cases}$

$y_i = w^T x_i + \varepsilon_i$ ε_i is a Gaussian random Variable with 0 mean.

7. 7.1 1. Initialize $w_0^0 = 0$

while not converged do

$w_0^{(t)} \leftarrow \sum_{i=1}^{N} (y_i - \hat{y}_i^{(t-1)}) / N + w_0^{(t-1)}$

for $k \in \{1, 2, \ldots d\}$ do

$\left.\begin{array}{l} \_\_\_\_\_\_ \\ \_\_\_\_\_\_ \\ \_\_\_\_\_\_ \end{array}\right\}$ same as Algorithm 1.

end

end

2. $O(\|X\|_0)$

3. $O(\|X\|_0 + 2)$, $O(N+2)$

4 $O(z_j)$

5. $\hat{y}_i^{(t+1)} = x_i w^{(t)} + w_0^{(t+1)}$

$= \hat{y}_i^{(t)} - w_0^{(t)} + w_0^{(t+1)}$

Complexity $O(N)$ or $O(z_j)$

6. $c_k^{(t+1)} = 2 \sum_{i=1}^{N} x_{ik} (y_i - \hat{y}_i + w_k^{(t)} x_{ik})$

$a_k^{(t+1)} = 2 \sum_{i=1}^{N} x_{ik}^2$

$$w_k^{(t+1)} = \begin{cases} (c_k^{(t+1)} + \lambda) / a_k^{(t+1)} & c_k < -\lambda \\ 0 & c_k \in [-\lambda, \lambda] \\ (c_k^{(t+1)} - \lambda) / a_k^{(t+1)} & c_k > +\lambda \end{cases}$$

$\hat{y}_i^{(t+1)} = \hat{y}_i^{(t)} + x_{ik} (w_k^{(t+1)} - w_k^{(t)})$

Complexity $O(N)$

7 $O(Nd)$

7 Programming Question 2: Lasso

7.3 Try out your work on synthetic data

Question 1

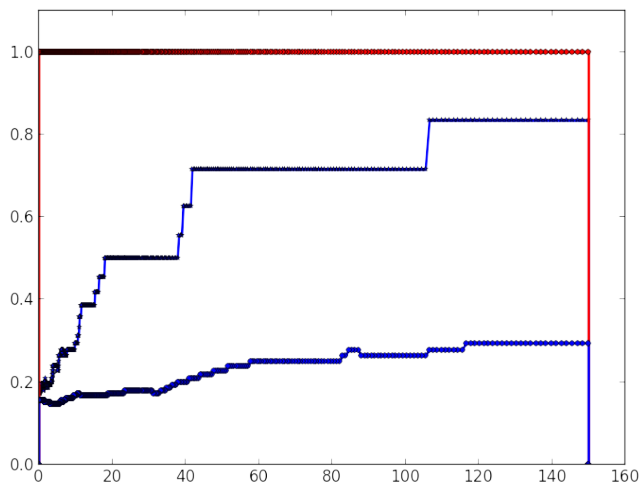


Figure 1: Recall(red) and precision(blue) plots of σ 1(star) and 10(circle) values against λ.

A large λ helps to more accurately discover non-zeros, because more false positive are discovered. As the λ goes lower, the precision goes down. But recalls are always 100%.

Question 2

Plots are found above with circle markers.
When I changed σ from 1.0 to 10.0, the precision become lower, because more false positive are discovered. This is caused by the noise introduced by the larger σ. But the recalls stay at 100%.
In this case, I'd like to start with a larger λ and decreased from there.

7.4 Become a data scientist at Yelp

Question 1

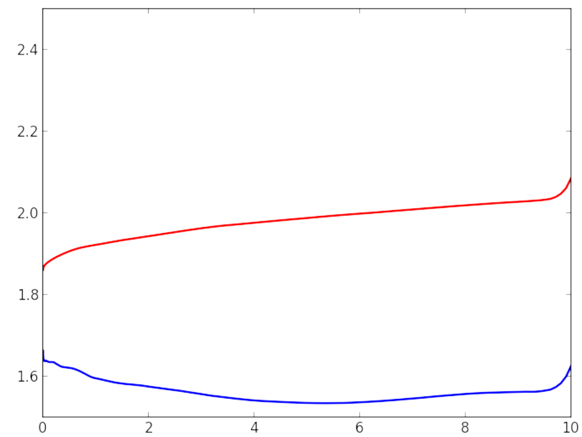


Figure 2: RMSE values against λ from training (red line) and testing (blue line) datasets.

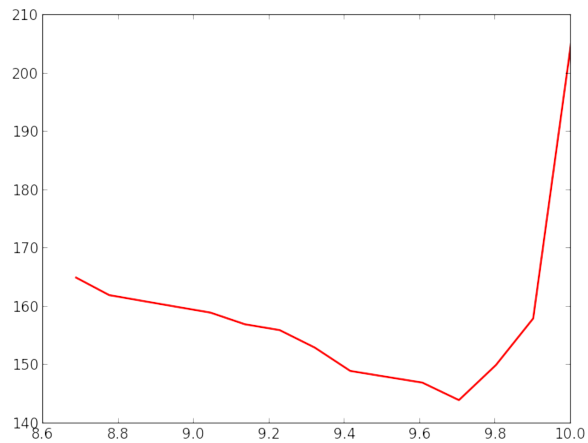


Figure 3: Number of nonzero weights against λ from training (red line) and testing (blue line) datasets.

Question 2

Using the best λ value 4.9983 from the validation performance, I got the RMSE value 1.1829717559 from the testing dataset.

Question 3

The best 10 features I discovered are :

sqrt(ReviewNumCharacters*UserCoolVotes) : [67.64008046]
sqrt(UserCoolVotes*BusinessNumStars) : [33.55206007]
sqrt(ReviewNumCharacters*UserFunnyVotes) : [14.92750312]
sqrt(UserFunnyVotes*InPhoenix) : [61.34628919]
BusinessNumReviews*InGlendale : [17.02279332]
ReviewInFall*InGlendale : [16.25769229]
UserUsefulVotes*InScottsdale : [29.18772702]
log(ReviewNumCharacters*UserUsefulVotes) : [35.83890522]
sq(UserUsefulVotes*IsRestaurant) : [42.39206274]
sqrt(ReviewNumWords*UserUsefulVotes) : [56.53814474]

Question 4

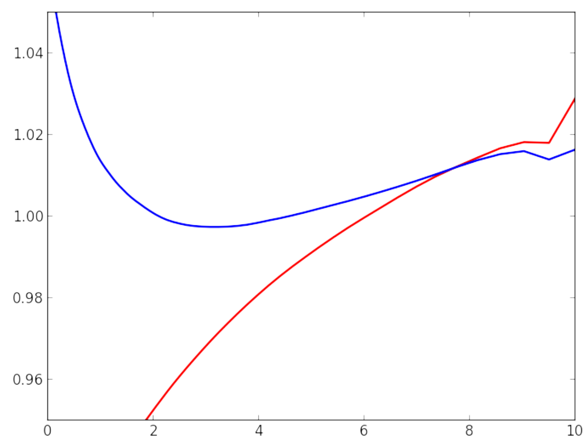


Figure 4: RMSE values against λ from training (red line) and testing (blue line) datasets.

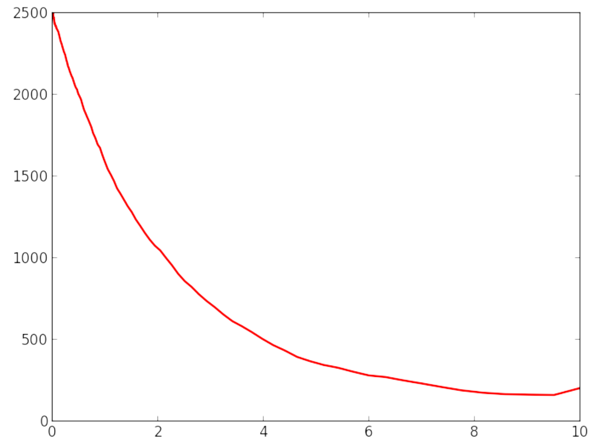


Figure 5: Number of nonzero weights against λ from training (red line) and testing (blue line) datasets.

Using the best λ value 3.0 from the validation performance, I got the RMSE value 0.5608 from the testing dataset.

The best 10 features I discovered are :
great : [21.10928191]
best : [18.00001071]
amazing : [14.74667601]
love : [11.76668926]
delicious : [11.09071936]
awesome : [12.56298077]
perfect : [9.47380946]
excellent : [8.90874206]
wonderful : [8.44758452]
friendly : [8.64004513]