Machine Learning - HW1

Faizaan Charania
11146 3646

Q1

We know that X_1 & X_2 are independent and uniformly distributed on $[0,1]$

$$X = \max(X_1, 2X_2)$$

Let $X_3 = 2X_2$ $\therefore$ X_3 is uniformly distributed on $[0,2]$

$\therefore$ $X = \max(X_1, X_3)$

- X performs differently in ranges $0 \leq x \leq 1$ and $1 \leq x \leq 2$

$\therefore$ for $\forall x, 0 \leq x \leq 1$

$$\begin{aligned} C.D.F(X) &= F_X(x) \\ &= Pr(X \leq x) \\ &= Pr(\max(X_1, X_3) \leq x) \\ &= Pr(X_1 \leq x, X_3 \leq x) \\ &= Pr(X_1 \leq x) \cdot Pr(X_3 \leq x) \quad [\because X_1 \perp X_3] \\ &= F_{X_1}(x) \cdot F_{X_3}(x) \end{aligned}$$

$$P.D.F(X) = \frac{\int}{dx} F_{X_1}(x) \cdot F_{X_3}(x)$$

$$= F_{X_1}(x) \cdot \int_{X_3}(x) + F_{X_3}(x) \cdot \int_{X_1}(x)$$

$$= \frac{x-0}{1-0} \cdot \frac{1}{2} + \frac{(x-0)}{(2-0)} \cdot 1$$

$$= \frac{x}{2} + \frac{x}{2} = x \;//$$

for $\forall x$, $1 < x \leq 2$

$$C.D.F(X) = P_R(X \leq x)$$

$$= P_R(\max(X_1, X_3) \leq x)$$

$$=$$

$\because X_1$ is not defined for $1 < x \leq 2$

~~$F_X(x) = P_R($~~

$$F_{X_1}(x) = 1 \qquad \forall x,\ 1 < x \leq 2$$

$$C.D.F(X) = P_R(X_1 \leq x) \cdot P_R(X_3 \leq x)$$

$$= 1 \cdot \frac{x}{2}$$

$$= \frac{x}{2}$$

$$\therefore P.D.F(X) = f_X(x) = \frac{\int}{dx} F_X(x) = \frac{\int}{dx} \frac{x}{2} = \frac{1}{2} \;//$$

1) $E(X)$

$$E[X] = \int_0^2 x\, f_x(x)\, dx$$

$$= \int_0^1 x f_x(x)\, dx + \int_1^2 x f_x(x)\, dx$$

$$= \int_0^1 x \times x\, dx + \int_1^2 x \times \frac{1}{2}\, dx$$

$$= \int_0^1 x^2\, dx + \int_1^2 \frac{x}{2}\, dx$$

$$= \left[\frac{x^3}{3}\right]_0^1 + \left[\frac{x^2}{4}\right]_1^2 = \frac{1}{3} + \frac{3}{4}$$

$$= \frac{13}{12} //$$

2) $Var(X)$

$$Var(X) = E[X^2] - (E[X])^2$$

$$E[X^2] = \int_0^2 x^2 f_x(x)\, dx$$

$$= \int_0^1 x^2 f_x(x)\, dx + \int_1^2 x^2 f_x(x)\, dx$$

$$= \int_0^1 x^3\,dx + \int_1^2 \frac{x^2}{2}\,dx$$

$$= \left[\frac{x^4}{4}\right]_0^1 + \left[\frac{x^3}{6}\right]_1^2$$

$$= \frac{1}{4} + \frac{7}{6} = \frac{17}{12}$$

$\therefore$ $\text{Var}(X) = E[X^2] - (E[X])^2$

$$= \frac{17}{12} - \left(\frac{13}{12}\right)^2 = \frac{17}{12} - \frac{169}{144}$$

$$= \frac{204 - 169}{144} = \frac{35}{144}$$ //

$\therefore$ $\text{Var}(X) = \frac{35}{144}$ //

Q2

2.1 Log-likelihood of X given λ

We know that X given λ is a Poisson distribution

$$\text{Likelihood}(X|\lambda) = P(X|\lambda)$$

$$= P(X_1 = x_1, X_2 = x_2 \cdots X_n = x_n \mid \lambda)$$

$\because$ All X_i are independent, we get

$$P(X|\lambda) = \prod_{i=1}^{n} P(X_i = x_i|\lambda)$$

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

$\therefore$ log-likelihood is

$$\log\left(\prod_{i=1}^{n} P(X_i = x_i|\lambda)\right) = \sum_{i=1}^{n} \log\left[\frac{\lambda^{x_i} e^{-\lambda}}{x_i!}\right]$$

$$= \sum_{i=1}^{n} (x_i \log\lambda - \log(x_i!) - \lambda)$$

$$= \sum_{i=1}^{n} (x_i \log\lambda - \log(x_i!)) - n\lambda$$

$\therefore$ Log-likelihood is $\sum_{i=1}^{n} (x_i \log\lambda - \log(x_i!)) - n\lambda$

2.2) M.L.E for λ in general case

$$M.L.E = \underset{\lambda}{\operatorname{argmax}} \sum_{i=1}^{n} \left(x_i \log\lambda - \log(x_i!) \right) - n\lambda$$

To find max we differentiate & equate to 0.

$$\therefore \frac{\delta}{\delta\lambda} \left(\Sigma(x_i \log\lambda - \log(x_i!)) - n\lambda \right) = 0$$

$$\sum_{i=1}^{n} x_i \frac{\delta}{\delta\lambda} \log\lambda - \frac{\delta}{\delta\lambda} \sum_{i=1}^{n} \log(x_i!) - n\frac{\delta}{\delta\lambda}\lambda = 0$$

$$\Rightarrow \frac{\sum_{i=1}^{n} x_i}{\hat{\lambda}} - 0 - n = 0$$

$$\therefore n = \sum_{i=1}^{n} x_i \times \frac{1}{\hat{\lambda}}$$

$$\therefore \boxed{\hat{\lambda} = \frac{\sum_{i=1}^{n} x_i}{n}}$$

2.1.3. To compute M.L.E for λ over observed X we can use previous solution 2.1.2

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

$$\therefore \hat{\lambda} = \frac{4+5+3+5+6+9+3}{7}$$

$$= \frac{35}{7} = 5 //$$

2.2.2 M.A.P

2.2.1 Given $\lambda \sim \Gamma(\alpha, \beta)$

λ is a Gamma distribution with a p.d.f

$$p(\lambda \mid \alpha, \beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}, \quad \lambda > 0$$

Posterior distribution over λ

$$P(\lambda \mid X) = \frac{P(X \mid \lambda) \cdot P(\lambda \mid \alpha, \beta)}{P(X)}$$

$$\therefore P(\lambda \mid X) \propto P(X \mid \lambda) \cdot P(\lambda \mid \alpha, \beta)$$

$$P(X \mid \lambda) \cdot P(\lambda \mid \alpha, \beta) = \left(\prod_{i=1}^{n} \frac{\lambda^{x_i}}{x_i!} e^{-\lambda} \right) \times \frac{\beta^{\alpha}}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}, \quad \lambda > 0$$

$$= \frac{\lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} (x_i!)} e^{-n\lambda} \times \frac{\beta^{\alpha}}{\Gamma(\alpha)} \cdot \lambda^{\alpha-1} e^{-\beta\lambda}$$

$$\therefore P(X|\lambda) \cdot P(\lambda|\alpha,\beta)$$

$$= \frac{\lambda^{\sum_{i=1}^{n} x_i + \alpha - 1}}{\prod_{i=1}^{n} x_i!} \times \frac{\beta^{\alpha}}{\Gamma(\alpha)} e^{-\lambda[\beta+n]}$$

$$\therefore P(\lambda|X) \propto P(X|\lambda)\, P(\lambda|\alpha,\beta)$$

$$P(\lambda|X) \propto \frac{\lambda^{\sum_{i=1}^{n} x_i + \alpha - 1}}{\prod_{i=1}^{n} x_i!} \times \frac{\beta^{\alpha}}{\Gamma(\alpha)} \cdot e^{-\lambda(\beta+n)}$$

222

Find Maximum a posterior (MAP) estimate of λ

To find M.A.P of λ, we take derivative of Posterior Distribution of λ w.r.t λ and then equate it to zero.

(Take log to make our calculation easy)

$$\therefore \frac{\delta}{\delta \lambda} \log\left(P(\lambda|x)\right) = 0$$

$$\log\left(P(\lambda|x)\right) = \log\left[\frac{\beta^{\alpha}}{\prod_{i=1}^{n}(x_i!)} \times \frac{\lambda^{\sum_{i=1}^{n} x_i + \alpha - 1}}{\Gamma(\alpha)} e^{-\lambda(\beta+n)}\right]$$

$$= \alpha \log \beta + \left(\sum_{i=1}^{n} x_i + \alpha - 1\right) \log \lambda - \lambda(\beta+n) -$$

$$\log\left(\prod_{i=1}^{n}(x_i!)\right) - \log\left(\Gamma(\alpha)\right)$$

$$\therefore \frac{\partial}{\partial \lambda}\left[\log P(\lambda|x)\right] =$$

$$\left[\sum_{i=1}^{n} x_i + \alpha - 1\right] \times \frac{1}{\lambda} - (\beta + n) = 0$$

$$\therefore \quad \hat{\lambda} = \frac{\left[\sum_{i=1}^{n} x_i + \alpha - 1\right]}{\beta + n}$$

2.3 Estimator Bias

2.3.1

Let $\hat{\eta} = e^{-2x}$

To prove,

$\therefore$ M.L.E of $\eta = \hat{\eta} = e^{-2x}$

$$\hat{\eta} = \underset{\eta}{\operatorname{argmax}}\, P(x|\eta)$$

$$\eta = e^{-2\lambda} \quad \text{(given)}$$

$$\therefore \log \eta = -2\lambda$$

$$\therefore \lambda = \left(\frac{-\log \eta}{2}\right) \quad \text{—(i)}$$

We know X is a Poisson s.t

$$X \sim \text{Poisson}(\lambda)$$

$$\therefore P(X|\lambda) = \frac{\lambda^k}{k!} e^{-\lambda} \quad \{k = 0, 1, 2 \ldots\} \quad \text{—(ii)}$$

From (i) & (ii) we get

$$P(X|\eta) = \frac{(\log \eta)^k}{(-2)^k k!} e^{-(\log \eta/-2)}$$

$$\therefore P(x|\eta) = \frac{(\log \eta)^k}{(-2)^k \lfloor k \rfloor_0} \eta^{1/2} \quad \text{---} \quad (iii)$$

Then to find out M.L.E (η) take derivative of (iii) w.r.t η & equate to zero

$$\frac{\delta}{\delta \eta}\left(P(x|\eta)\right) = \frac{k(\log \eta)^{k-1}}{(-2)^k \lfloor k \rfloor_0}\left(\frac{1}{\eta}\right) \times \eta^{1/2} +$$

$$\frac{(\log \eta)^k}{(-2)^k \lfloor k \rfloor_0} \times \frac{1}{2} \times \eta^{-1/2}$$

$$= \frac{1}{(-2)^k \lfloor k \rfloor_0} \times (\log \eta)^{k-1}\left[\frac{k}{\eta^{1/2}} + \frac{(\log \eta)^k}{2\eta^{1/2}}\right] = 0$$

$$\therefore \quad \frac{k}{\eta^{1/2}} + \frac{(\log \eta)^k}{2\eta^{1/2}} = 0$$

$$\therefore (\log \eta)^k = -k$$

$$\hat{\eta} = e^{-k}$$

∴ we only have one observation of X

$$\boxed{\hat{\eta} = e^{-2x}}$$

2.3.2 $\text{Bias}(\hat{\eta}) = E[\hat{\eta}] - \eta$

$$\eta = e^{-2\lambda}$$

$$E[\hat{\eta}] = \sum_{x \geq 0} \hat{\eta} P(x)$$

$$= \sum_{x \geq 0} e^{-2x} \frac{\lambda^x}{x!} e^{-\lambda}$$

$$= e^{-\lambda} \sum_{x \geq 0} \frac{e^{-2x} \lambda^x}{x!}$$

$$= e^{-\lambda} \sum_{x=0}^{\infty} \frac{(\lambda e^{-2})^x}{x!}$$

Using the Taylor series

$$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}$$

$$\therefore E[\hat{\eta}] = e^{-\lambda} \cdot e^{\lambda e^{-2}}$$

$$= e^{\lambda(1/e^2 - 1)}$$

$$\therefore \text{bias}(\hat{\eta}) = E[\hat{\eta}] - \eta$$

$$= e^{\lambda(1/e^2 - 1)} - e^{-2\lambda}$$

$\because$ bias$(\hat{\eta})$ is just a value, we can also say bias$(\hat{\eta}) = e^{-2\lambda} - e^{\lambda(1/e^2 - 1)}$

3.3

We assume $\hat{\eta} = (-1)^x$

∴ To find out the expected value, we do

$$E[\hat{\eta}] = \sum_{x=0}^{\infty} \hat{\eta} P(x)$$

$$\hat{\eta} = (-1)^x$$

$$\therefore E[\hat{\eta}] = \sum_{x=0}^{\infty} (-1)^x \frac{\lambda^x e^{-\lambda}}{x!}$$

$$= e^{-\lambda} \sum_{x=0}^{\infty} \frac{(-\lambda)^x}{x!}$$

Using the taylor series, $e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}$

$$\therefore \sum_{x=0}^{\infty} \frac{(-\lambda)^x}{x!} = e^{-\lambda}$$

$$\therefore E[\hat{\eta}] = e^{-\lambda} * e^{-\lambda} = e^{-2\lambda}$$

$$\text{bias}(\hat{\eta}) = E[\hat{\eta}] - \eta$$

$$= e^{-2\lambda} - e^{-2\lambda}$$

$$= 0$$

∴ $(-1)^x$ is an unbiased estimator

$\hat{\eta} = (-1)^x$ totally depends on whether X is even or odd.

For all even X, η is positive & η is negative for odd values of X.

∴ It is a bad estimator to use.

Q3 Regression and M.L.E

$y_i = w^T x_i + \epsilon_i$ is generated by a linear regression model.

3.1 Assumption $\epsilon_1, \ldots \epsilon_n$ are independent each with mean 0 & variance σ_i^2

$\therefore \epsilon_i \sim N(0, \sigma_i^2)$

$$P(Y|w,\sigma) = P(Y=y_1|w,\sigma) \cdot P(Y=y_2|w,\sigma) \cdots P(Y=y_n|w,\sigma)$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(\frac{-1}{2\sigma_i^2}(y_i - w^T x_i)^2\right)$$

$\therefore$ Log - liklihood is

$$\log(P(Y|w,\sigma)) = \log\left(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(\frac{-1}{2\sigma_i^2}(y_i - w^T x_i)^2\right)\right)$$

$$= \sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi}\,\sigma_i}\right) - \sum_{i=1}^{n} \left(\frac{1}{2\sigma_i^2}\left(y_i - w^T x_i\right)^2\right)$$

$$\underbrace{\qquad\qquad}_{a} \qquad\qquad \underbrace{\qquad\qquad\qquad\qquad}_{b}$$

$\therefore$ a is independent of w, to find max $P(Y \mid w, \sigma)$ we ~~minimize~~ find min b

$\therefore$ M.L.E

$$\hat{w} = \underset{w}{\operatorname{argmin}} \sum_{i=1}^{n} \frac{1}{2\sigma_i} \left(y_i - w^T x_i\right)^2$$

we can put $\frac{1}{2\sigma_i} = s_i$

$$\therefore \hat{w} = \underset{w}{\operatorname{argmin}} \sum_{i=1}^{n} s_i \left(y_i - w^T x_i\right)^2$$

$\therefore$ M.L.E can be found by minimising the above function.

3.2

Likelihood is still $P(Y|\omega, \epsilon)$

$$P(Y|\omega, \epsilon) = \prod_{i=1}^{n} \frac{1}{2b} \exp\left(-\frac{|y_i - \omega^T x_i|}{b}\right)$$

$\because \epsilon \sim \text{Laplace}(b)$

Taking log on both sides

$$\log(P(Y|\omega, \epsilon)) = \log\left(\prod_{i=1}^{n} \frac{1}{2b} \exp\left(-\frac{|y_i - \omega^T x_i|}{b}\right)\right)$$

$$= \sum_{i=1}^{n} \left[\log \frac{1}{2b} - \frac{|y_i - \omega^T x_i|}{b}\right]$$

$$= \underbrace{\sum_{i=1}^{n} \log \frac{1}{2b}}_{a} - \underbrace{\frac{1}{b} \sum_{i=1}^{n} |y_i - \omega^T x_i|}_{b}$$

As we have seen in Q 3.1 a is independent of ω.

$\therefore$ As done before

$$\text{M.L.E of } \omega = \hat{\omega} = \underset{\omega}{\text{argmin}} \frac{1}{b} \sum_{i=1}^{n} |y_i - \omega^T x_i|$$

$$= \underset{\omega}{\text{argmin}} \sum_{i=1}^{n} |y_i - \omega^T x_i|$$

$$\therefore \hat{\omega} = \underset{\omega}{\operatorname{argmin}} \left| y_i - \omega^T x_i \right|$$

3.3

Given the model, the residual is

$$r = \sum_{i=1}^{n} y_i - (\omega^T x_i + \epsilon_i)$$

If we assume ϵ_i to be from a Gaussian distribution $N(0, \sigma_i^2)$

the residual is depended on the square of parameter σ.

If we assume the noise to be a Laplace distribution, then it is only linearly depended on the value of b.

$\therefore$ In case of noise, the model with Laplace noise will be more robust.

Q4 Programming

4.1

4.1.1 Residual $r_i = y_i - (w^T x_i + b_i)$

∴ In Matrix form

$$R = Y - (W^T X + b)$$

where $R \rightarrow \mathbb{R}^{n \times 1}$

$Y \rightarrow \mathbb{R}^{n \times 1}$

$W \rightarrow \mathbb{R}^{d \times 1}$

$X \rightarrow \mathbb{R}^{d \times n}$

$b \rightarrow \mathbb{R}^{n \times 1}$

Time complexity $= O(\|X\|_0)$

∵ Dot product depends on number of non-zero entries in X.

4.1.2 Update rule for b

$$b = \frac{1}{n}\left[\vec{1} \cdot R\right] + b$$

where $\vec{1} \Rightarrow$ a ones-vector of length n

Time complexity : $O(n)$

$\because O(n)$ is required to calculate the vector dot product.

4.1.3 To update R we need to update the bias term

$$\therefore R = R + b_{old} - b$$

$b_{old} \rightarrow$ 'b' before update

$b \rightarrow$ 'b' after update

Time complexity : $O(n)$

$\because$ we need to update n values with $O(1)$

Update rule for C_k using $\hat{r}$

$$C_k \leftarrow 2 \sum_{i=1}^{n} x_{ik} (\hat{r}_i + w_k x_{ik})$$

∴ In Matrix form,

$$C_k \leftarrow 2x \left(x_k \cdot (\hat{r} + (x_k \cdot w_k))^T \right)$$

$x_k \cdot w_k$ takes z_k time

but $\hat{r}$ has 'n' entries, ∴ addition takes $O(n)$

∴ Time complexity $O(n)$

v)

$$\hat{r} \leftarrow \hat{r} + [w_k old - w_k new] x_k$$

Time complexity $O(z_k)$

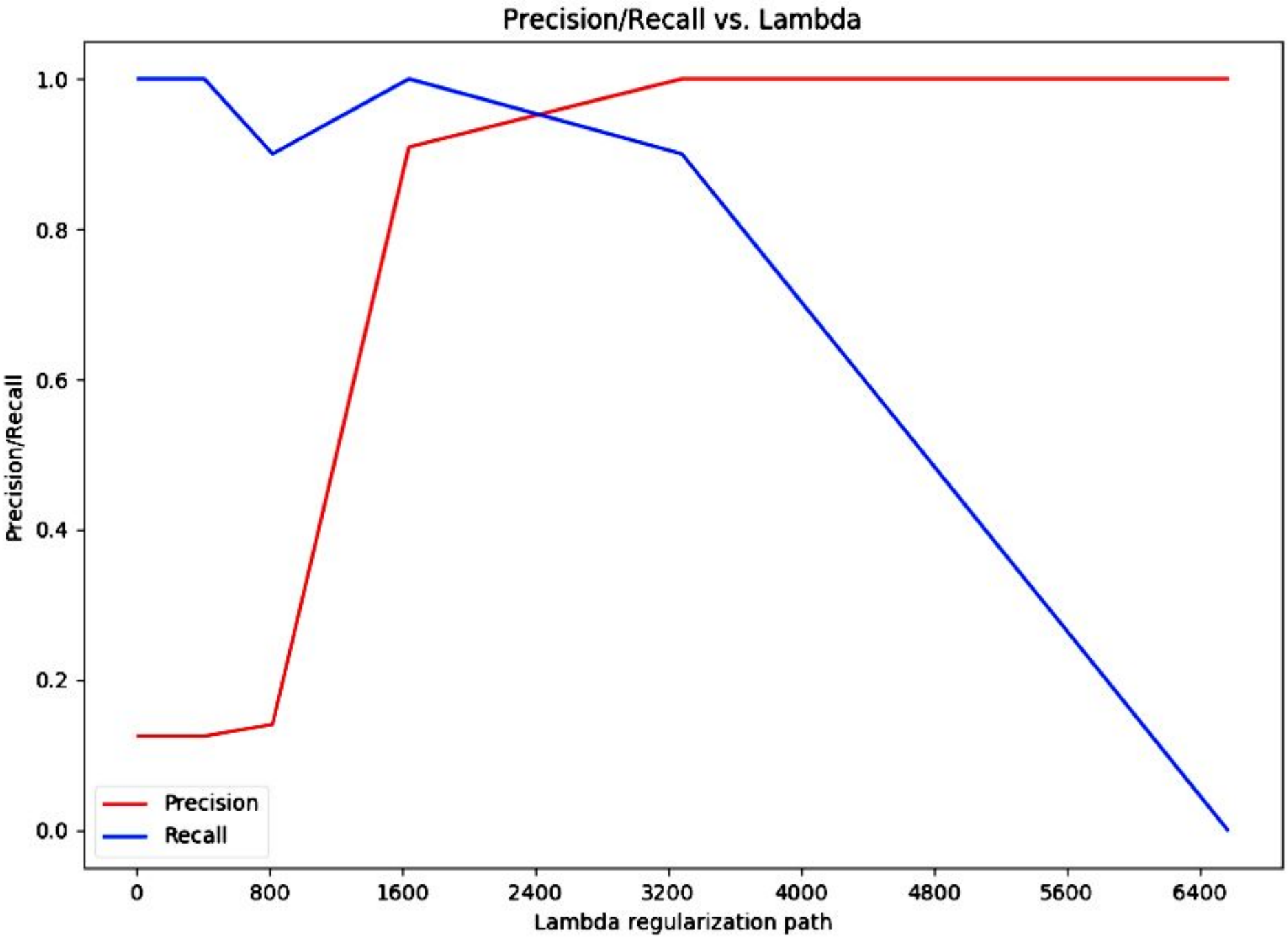
Per iteration complexity

$$O(\max(\max(n, \|X\|_0), nd))$$

Q 4.3.

4.3.1 The discovery of true non-zeros depends on the value of λ

For a higher λ precision is high while recall is low. As lambda decreases precision decreases & recall increases.


Precision/Recall vs. Lambda
Precision/Recall
1.0
0.8
0.6
0.4
0.2
0.0
0
800
1600
2400
3200
4000
4800
5600
6400
Lambda regularization path
Precision
Recall

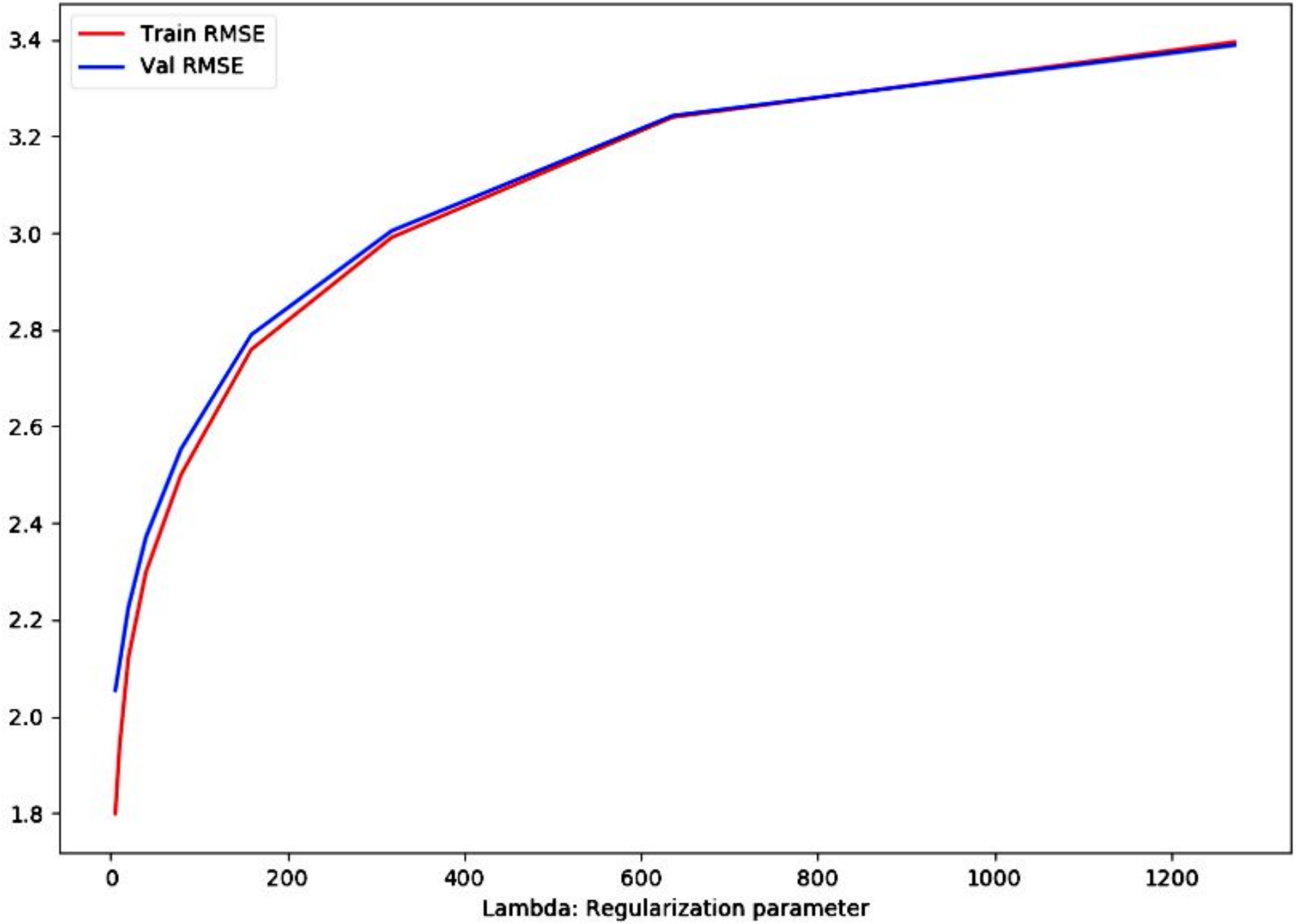
4.3.2

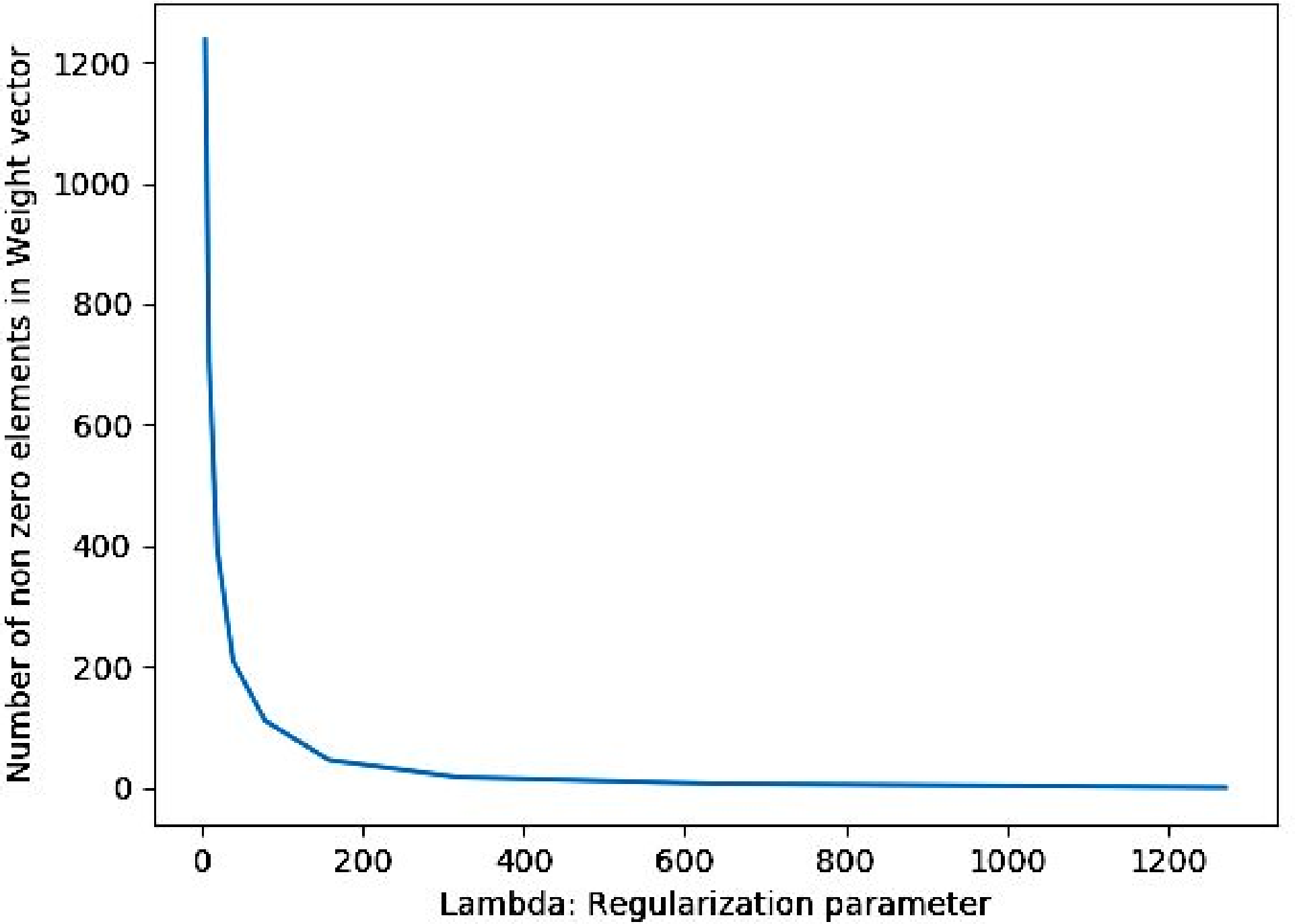
We use $\lambda = 1626.8$, which gave precision = 0.909 & recall = 1.0

when used for $\sigma = 10$, precision dropped to 0.1667 while recall stayed at 1

If we increase λ, we will get a higher precision.

4.4.1 Plots.


Train RMSE
Val RMSE
3.4
3.2
3.0
2.8
2.6
2.4
2.2
2.0
1.8
0
200
400
600
800
1000
1200
Lambda: Regularization parameter


Number of non zero elements in Weight vector
1200
1000
800
600
400
200
0
0
200
400
600
800
1000
1200
Lambda: Regularization parameter

1.2

The λ that achieves the best performance was

$$\lambda = 2.484 \quad , \quad R.M.S.E = 2.048$$

Top 10 features were

- Stars + 10.93
- Spearmint + 10.17
- big 9.11
- lifesaver 9.09
- truly 8.56
- sweet black 8.50
- nearly 8.27
- acidity provides 8.22
- lemony 8.02
- agreeability 7.89

Bottom 10 features

- earns -9.16
- cherry berry -8.49
- black -8.04
- soft -7.49
- spices -7.08
- sparkler -6.96
- brightened -6.82
- low alcohol -6.77
- cuts -6.71
- high -6.54

While some features are intuitive, others might have very high / low weight because they occur with important features.

submission.csv 2.05024

4 hours ago by FaizaanCharania

add submission details