CMPUT466 AS2

Bowei Wang
1462495

CMPUT 466 AS 2

Bowei Wang
1402495

Question 1.

(a) We have the Bayes rule: $p(f|D) = \dfrac{P(D|f) \cdot p(f)}{P(D)}$

because $P(D)$ does not affect the result, we can simplify it to: $\propto P(D|f) \cdot P(f)$

Therefore we just need to solve: $f_{MAP} = \underset{f \in F}{\operatorname{argmax}} \{P(D|f) P(f)\}$

Besides, we need to know the log-likelihood function:

$P(D|\lambda)$ for gaussian dist:

$$\ln(D|\theta, \sigma_0^2) = \ln \prod_{i=1} p(x_i|\theta, \sigma_0^2) = n \ln \frac{1}{\sqrt{2\pi}} + n \ln \frac{1}{\sigma_0} - \frac{\sum_{i=1}^{n} (x_i - \theta)^2}{2\sigma_0^2}$$

Besides, we need to know the probability density function of prior, which is also a Gaussian dist:

$$N(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

The prior distribution is: $p(\lambda) = \dfrac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(\lambda-\mu)^2}{2\sigma^2}}$

now, we just need to maximize the logrithm of the posterior distribution $p(\lambda|D)$, by using:

$$\ln p(\lambda|D) \propto \ln p(D|\lambda) + \ln p(\lambda)$$

$$= n \ln \frac{1}{\sqrt{2\pi}} + n \ln \frac{1}{\sigma_0} - \frac{\sum_{i=1}^{n} (x_i - \lambda)^2}{2\sigma_0^2} - \frac{(\lambda-\mu)^2}{2\sigma^2} \cdot \frac{1}{\sqrt{2\pi\sigma^2}}$$

take the first derivative: $\dfrac{\sum_{i=1}^{n} (x_i - \lambda)}{\sigma_0^2} - \dfrac{(\lambda - \mu)}{\sigma^3 \sqrt{2\pi}} = 0$

that is $\dfrac{\sum_{i=1}^{n} (x_i - \theta)}{\sigma_0^2} - \dfrac{(\theta - \mu)}{\sigma^3 \sqrt{2\pi}} = 0$

(b) We have the same log likelihood function as question a

so $\ln P(D|\lambda) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\lambda)^2$.

Besides, we know the pdf of laplace distribution:

$P(x) = \frac{1}{2b}\exp\left(\frac{-|x-\mu|}{b}\right)$

Therefore, the prior distribution is:

$P(\lambda) = \frac{1}{2b}\exp\left(\frac{-|\lambda-\mu|}{b}\right)$

Now, we can try to maximize the logorithm of the posterior distribution $P(\lambda|D)$ using:

$\ln P(\lambda|D) \propto \ln p(D|\lambda) + \ln p(\lambda)$

$= -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\lambda)^2$

$+ \frac{1}{2b}\cdot\left(\frac{-|\lambda-\mu|}{b}\right)$

While, we don't know wheather λ is larger than 0 or less than 0 (μ is 0), we need to consider them seperately

$$\theta_{MAP} \text{ can be either } \begin{cases} -\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\theta) - \frac{1}{2b^2} & (\text{when } \theta \geq 0) \\ \text{or} \\ -\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\theta) + \frac{1}{2b^2} & (\text{when } \theta < 0) \end{cases}$$

We can solve this equation with value passed in. by seperating the posterior estimate into two different cases.

(c)

Now that we are handling with the multivariance Gaussian distribution. We have:

$$p(w) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left(-\frac{1}{2}(w-\mu)^T \Sigma^{-1} (w-\mu)\right),$$

In this question, we have the prior of the θ: from the Gaussian dist: $N(\mu = 0, \Sigma = \sigma^2 I)$

$$P(\theta) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left(-\frac{1}{2}(\theta-\mu)^T \Sigma^{-1} (\theta-\mu)\right).$$

with $\mu = 0$, $\Sigma = \sigma^2 I$.

$$p(\theta) = \frac{1}{\sqrt{(2\pi)^d |\sigma^2 I|}} \exp\left(-\frac{1}{2}(\theta)^T (\sigma^2 I)^{-1} \cdot \theta\right)$$

Besides, the log likelihood function for multivariance Gaussian distribution is:

$$-\frac{1}{2}\left(\ln|\Sigma|) + (x-\mu)^T \Sigma^{-1} (x-\mu) + k\ln(2\pi)\right)$$

Given a set X of iid, vectors, we have:

$$\log N = -\frac{n}{2}\ln|\Sigma_0| - \frac{1}{2}\sum_{i=1}^{n}(x_i - \theta)^T \Sigma_0^{-1} (x_i - \theta)$$

We know that $\lambda_{MAP} = \underset{\lambda \in (0,\infty)}{\operatorname{argmax}} \{P(D|\lambda)P(\lambda)\}$

we just need to consider:

$$\ln(\theta|D) \propto \ln p(D|\theta) + \ln p(\theta)$$

$$= -\frac{n}{2}(\ln|\Sigma_0|) - \frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^T\Sigma_0^{-1}(x_i-\theta)$$

$$+ \frac{-\frac{1}{2}\theta^T(\sigma^2)^{-1}\cdot\theta}{\sqrt{(2\pi)^d\cdot(\sigma^2)}}$$

We have the rule that: $\frac{d(x^T a)}{dx} = \frac{d(a^T x)}{dx} = a^T$

Therefore, take the first derivative:

$$-\frac{1}{2}\sum_{i=1}^{n}((\Sigma_0^{-1})^T(x_i-\theta) + \Sigma_0^{-1}(x_i-\theta)^T)$$

$$-\frac{1}{2}\frac{(((\sigma^2)^{-1})^T\cdot\theta + \theta^T(\sigma^2)^{-1})}{\sqrt{(2\pi)^d\cdot(\sigma^2)}} = 0$$

Therefore, we can get the MAP estimate by put in the value and calculate the real value of θ.

cmput466 assignment2 question2

(a):

I have kept increasing the number of features and I found when the number of features comes to 80, it does not work anymore.

An error happened. It says :

"   raise LinAlgError("Singular matrix")   numpy.linalg.linalg.LinAlgError: Singular matrix"

It does not work for any feature number in between 80-385.

(b)standard error is reported

(c)Ridge Regression is added.

Different from feature select linear regression in (a), even all the features are included, there is no error.

The reason is that when the feature number becomes larger, the product of XT •X could be sigular matrix and cannot be inversed.

However in ridge regression, an identity matrix * lambda will be added to XT • X that aviod the sigular case.

(d)Lasso is added

(e)SGD is added, the Average error for SGD is : 0.24847505429459588, standard error for SGD is : 0.00012867571444407378

(f)batch gradient descent added

The error decreases for batch gradient and slightly increases for stochastic gradient descent as the epoch increase, but the standard error for both are decrease.

Besides, as you increase the number of epoch, the running time also increases, that means for batch gradient descent, the error decreases as the running time becomes longer.