CSE/ISYE 6740 Homework 3 (Le Song)

Siying Cen (GTID: scen9)

deadline: 11/14

1 Linear Regression

(a) Using the normal equation, and the model (Eqn.1), derive the expectation $E[\hat{\theta}]$. Note that here X is fixed, and only Y is random, i.e. "fixed design" as in statistics.

Answer: According to the definition of $\hat{\theta}$:

$$\begin{aligned}(\hat{\theta}) &= E((X^TX)^{-1}X^TY)) \\ &= E((X^TX)^{-1}X^T(X\theta+\epsilon))) \\ &= (X^TX)^{-1}X^TE(X\theta+\epsilon)) \\ &= (X^TX)^{-1}X^T(XE(I\theta)+E(\epsilon))\end{aligned}$$

Since $E(\epsilon)=0$ and $E(\theta)=\theta$

$$E(\hat{\theta})=\theta$$

(b) Similarly, derive the variance $\mathrm{V}_{ar}[\hat{\theta}]$.

Answer:

$$\begin{aligned}\mathrm{V}_{ar}(\hat{\theta}) &= \mathrm{V}_{ar}((X^TX)^{-1}X^TY) \\ &= (X^TX)^{-1}X^T\mathrm{V}_{ar}(X\theta+\epsilon)((X^TX)^{-1}X^T)^T \\ &= (X^TX)^{-1}X^T\mathrm{V}_{ar}(\epsilon)((X^TX)^{-1}X^T)^T \\ &= (X^TX)^{-1}X^T\sigma^2I((X^TX)^{-1}X^T)^T \\ &= \sigma^2I(X^TX)^{-1}X^TX((X^TX)^{-1})^T \\ &= \sigma^2I(X^TX)^{-1}\end{aligned}$$

(c) Under the white noise assumption above, someone claims that $\hat{\theta}$ follows Gaussian distribution with mean and variance in (a) and (b), respectively. Do you agree with this claim? Why or why not?

Answer: Yes, $\hat{\theta}$ follows Gaussian distribution because $\hat{\theta}=(X^TX)^{-1}X^T(X\theta+\epsilon))$, where θ is constant, X is fixed, and ϵ^i follows Gaussian distribution. Hence, $\hat{\theta}$ also follows Gaussian distribution.

(d) Weighted linear regression

Answer: Recall Eqn.(1), the Y^i should follow $\mathcal{N}(0,\sigma_i^2 I)$. Therefore, the probabilistic expression:

$$p(y^i|x^i;\theta)=\frac{1}{\sqrt{2\pi}\sigma_i}\exp\left(-\frac{(\theta^Tx^i-y^i)^2}{2\sigma_i^2}\right)$$

Based on independence assumption, the likelihood:

$$L(\theta)=\prod_{i=1}^n p(y^i|x^i;\theta)=\prod_{i=1}^n\frac{1}{\sqrt{2\pi}\sigma_i}\exp\left(-\frac{(\theta^Tx^i-y^i)^2}{2\sigma_i^2}\right)$$

Hence:

$$\log L(\theta) = \sum_{i=1}^{n} \left[\log \frac{1}{\sqrt{2\pi}\sigma_i} - \frac{(\theta^T x^i - y^i)^2}{2\sigma_i^2} \right]$$

Maximizing the likelihood:

$$\frac{\partial \log L(\theta)}{\partial \theta} = \sum_{i=1}^{n} - \frac{(\theta^T x^i - y^i)x^i}{\sigma_i^2} = 0$$

which is equivalent to:

$$-\sum_{i=1}^{n} \frac{x^i {x^i}^T \theta}{\sigma_i^2} + \sum_{i=1}^{n} \frac{y^i x^i}{\sigma_i^2} = 0$$

Therefore:

$$\hat{\theta} = (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} Y$$

where $X = (x^1, x^2, ..., x^n)$, and $Y = (y^1, y^2, ..., y^n)^T$

2 Ridge Regression

Show that the ridge regression estimate is the mean of the posterior distribution under a Gaussian prior $\theta \sim \mathcal{N}(X\theta, \sigma^2 I)$. Find the explicit relation between the regularization parameter λ in the ridge regression estimate of the parameter θ, and the variances σ^2,τ^2.

Answer: The posterior probability:

$$\begin{aligned}
p(\theta|y) &= \frac{p(y|\theta)p(\theta)}{p(y)} \\
&= \frac{\mathcal{N}(y|\sigma^2)\mathcal{N}(\theta|\tau^2)}{\Sigma' \mathcal{N}(\theta|\tau^2)\mathcal{N}(y|\sigma^2)} \\
&= \frac{1}{C}\mathcal{N}(y|X\theta, \sigma^2 I)\mathcal{N}(\theta|0, \tau^2) \\
&= \frac{1}{C} \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\left(- \frac{(y - X\theta)^T (y - X\theta)}{2\sigma^2} \right) \frac{1}{(\sqrt{2\pi}\tau)^n} \exp\left(- \frac{\theta^T \theta}{2\tau^2} \right) \\
&= \frac{1}{C(\sqrt{2\pi}\sigma)^n (\sqrt{2\pi}\tau)^n} \exp\left(- \frac{(y - X\theta)^T (y - X\theta)}{2\sigma^2} - \frac{\theta^T \theta}{2\tau^2} \right)
\end{aligned}$$

Therefore,

$$\frac{\partial \log p(\theta|y)}{\partial \theta} = \frac{2X^T (y - X\theta)}{2\sigma^2} - \frac{2\theta}{2\tau^2} = 0$$

$$\Rightarrow \hat{\theta} = \left(X^T X + \frac{\sigma^2}{\tau^2 I} \right)^{-1} X^T y$$

Compared with the ridge regression estimation $(X^T X + \lambda I)^{-1} X^T y$, we have $\lambda = \frac{\sigma^2}{\tau^2}$ Consider the posterior follows Gaussian distribution, we can define $p(\theta|y)$ $\mathcal{N}(\mu, \Sigma)$, whose exponent part of Gaussian equation is:

$$-\frac{1}{2}(\theta - \mu)^T \Sigma^{-1} (\theta - \mu) = -\frac{1}{2}(\theta^T \Sigma^{-1} \theta - 2\theta^T \Sigma^{-1} \mu + \theta^T \mu^{-1} \mu) = - \frac{(y - X\theta)^T (y - X\theta)}{2\sigma^2} - \frac{\theta^T \theta}{2\tau^2}$$

where the second order of θ:

$$\theta^T \Sigma^{-1} \theta = \frac{1}{\sigma^2} \theta^T X^T X \theta + \frac{1}{\tau^2} \theta^T \theta = \frac{1}{\sigma^2} \theta^T (X^T X + \frac{\sigma^2}{\tau^2} I) \theta$$

$$\Rightarrow \Sigma^{-1} = \frac{1}{\sigma^2} (X^T X + \frac{\sigma^2}{\tau^2} I)$$

and the first order of θ with μ:

$$\theta^T \sigma^{-1} \mu = \frac{1}{\sigma^2} \theta^T X^T y$$

$$\Rightarrow \mu = (X^T X + \frac{\sigma^2}{\tau^2} I) X^T y$$

3 Bayes Classifier

3.1 Bayes Classifier With General Loss Function

Write down the Bayes classifier $f : X \to Y$ for binary classification $Y \in \{-1, +1\}$. Simplify the classification rule as much as you can.

Answer: The loss function for this problem could be written as:

$$\begin{cases} 0, \text{ if } Y_1 = Y_2 = 1 \text{ or } Y_1 = Y_2 = -1 \\ p, \text{ if } Y_1 = 1, Y_2 = -1 \\ q, \text{ if } Y_1 = -1, Y_2 = 1 \end{cases}$$

And

$$P(Y = i|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

Therefore,

$$\begin{aligned} L_1 &= P(Y_1 = 1|X)L(Y_1 = 1, Y_2 = 1) + P(Y_1 = -1|X)L(Y_1 = -1, Y_2 = 1) \\ &= P(Y_1 = -1|X)p \end{aligned}$$

$$\begin{aligned} L_{-1} &= P(Y_1 = 1|X)L(Y_1 = 1, Y_2 = -1) + P(Y_1 = -1|X)L(Y_1 = -1, Y_2 = -1) \\ &= P(Y_1 = 1|X)q \end{aligned}$$

The Bayes decision rule:

$$\frac{L_1(X)}{L_{-1}(X)} = \frac{P(Y_1 = -1|X)p}{P(Y_1 = 1|X)q} = \frac{P(Y_1 = -1P(X|Y_1 = -1))p}{P(Y_1 = 1)P(X|Y_1 = 1)q}$$

Hence the classification rule:

$$f(X) = \begin{cases} 1, \text{ if } P(Y_1 = -1|X)p > P(Y_1 = 1|X)q \\ -1, \text{ if } P(Y_1 = -1|X)p < P(Y_1 = 1|X)q \end{cases}$$

3.2 Gaussian Class Conditional Distribution

(a) Based on the general loss function in problem 3.1, write the Bayes classifier as $f(X) = sign(h(X))$ and simplify h as much as possible. What is the geometric shape of the decision boundary:

Answer: In this case,

$$P(X|Y = i) = P(X|\mu_i, \Sigma_i) = \frac{1}{\sqrt{2\pi^D |\Sigma_i|}} \exp(-(X - \mu_i)^T \Sigma_i^{-1} (X - \mu_i)/2)$$

Define $f(X) = \text{sign}(\log(g(X)))$, then

$$\begin{aligned} g(X) &= \frac{P(Y_1 = 1)P(X|Y_1 = 1)p}{P(Y_1 = -1)P(X|Y_1 = -1)q} \\ &= \frac{P(Y_1 = 1)P(X|\mu_1, \Sigma_1)p}{P(Y_1 = -1)P(X|\mu_{-1}, \Sigma_{-1})q} \\ &= \frac{P(Y_1 = 1)\Sigma_{-1}^{-1/2}}{P(Y_1 = -1)\Sigma_1^{-1/2}} \exp\left(\frac{-(X - \mu_1)^T \Sigma_1^{-1} (X - \mu_1) + (X - \mu_{-1})^T \Sigma_{-1}^{-1} (X - \mu_{-1})}{2}\right) \end{aligned}$$

Thus

$$\begin{aligned} h(X) &= \log\left(\frac{P(Y_1 = 1)\Sigma_{-1}^{-1/2}}{P(Y_1 = -1)\Sigma_1^{-1/2}}\right) - 0.5((X - \mu_1)^T \Sigma_1^{-1} (X - \mu_1) \\ &- (X - \mu_{-1})^T \Sigma_{-1}^{-1} (X - \mu_{-1})) \\ &= \text{constant} - 0.5(X^T(\Sigma_1^{-1} - \Sigma_{-1}^{-1})X - (\mu_1^T \Sigma_{-1}^{-1} - \mu_{-1}^T \Sigma_{-1}^{-1})X \\ &- X^T(\mu_1 \Sigma_1^{-1} - \mu_{-1}\Sigma_{-1}^{-1}) - \mu_1^t \Sigma_1^{-1} \mu_1 + \mu_{-1}^t \Sigma_{-1}^{-1} \mu_{-1}) \end{aligned}$$

which is a n-dimensional quadratic surface when $\Sigma_1^{-1} \neq \Sigma_{-1}^{-1}$.

(b) Assume the two Gaussians have identical covariance matrices, repeat (a).

Answer: When $\Sigma_1^{-1} = \Sigma_{-1}^{-1}$,

$$h(X) = \text{constant} + 0.5((\mu_1^T - \mu_{-1}^T)\Sigma_1^{-1}X + X^T(\mu_1 - \mu_{-1})\Sigma_1^{-1} + \mu_1^t\Sigma_1^{-1}\mu_1 - \mu_{-1}^t\Sigma_1^{-1}\mu_{-1})$$

which is a n-dimensional plane.

(c) Assume the two Gaussians have covariance matrix which is equal to the identity matrix, repeat (a).

Answer: When $\Sigma_1^{-1} = \Sigma_{-1}^{-1} = I$,

$$\begin{aligned} h(X) &= \text{constant} + 0.5((\mu_1^T - \mu_{-1}^T)X + X^T(\mu_1 - \mu_{-1})) \\ &= \text{constant} + (\mu_1^T - \mu_{-1}^T)X \end{aligned}$$

which is a n-dimensional plane orthogonal to $\mu_1^T - \mu_{-1}^T$.

4 Logistic Regression

(a) Show that log-odds of success is a linear function of X.

Answer:

$$P[Y = 0|X = x] = 1 - P[Y = 1|X = x] = \frac{1}{1 + \exp(\omega_0 + \omega^T x)}$$

Therefore, the log-odds of success is

$$\ln\left(\frac{P[Y = 1|X = x]}{P[Y = 0|X = x]}\right) = \ln(\exp(\omega_0 + \omega^T x)) = \omega_0 + \omega^T x$$

which is a linear function of X.

(b) Show that the logistic loss $L(z) = \log(1 + \exp(-z))$ is a convex function.

Answer:

$$\frac{d^2L}{dz^2} = \frac{d}{dz}\left(\frac{-\exp(-z)}{1 - \exp(-z)}\right) = \frac{\exp(-z)}{(1 + \exp(-z))^2} > 0$$

Q.E.D.

5 Programming: Recommendation System

(a) Derive the update formula in (6) by solving the partial derivative.

Answer:

$$\begin{aligned} \frac{\partial E(U,V)}{\partial U_{v,k}} &= \frac{\partial(M_{v,j} - \Sigma_{k=1}^r U_{v,k}V_{j,k})^2}{\partial U_{v,k}} \\ &= -2V_{j,k}(\partial(M_{v,j} - \Sigma_{k=1}^r U_{v,k}V_{j,k}) \end{aligned}$$

$$\begin{aligned} \frac{\partial E(U,V)}{\partial V_{v,k}} &= \frac{\partial(M_{v,j} - \Sigma_{k=1}^r U_{v,k}V_{j,k})^2}{\partial V_{v,k}} \\ &= -2U_{j,k}(\partial(M_{v,j} - \Sigma_{k=1}^r U_{v,k}V_{j,k}) \end{aligned}$$

thus, formula in (6) can be updated to :

$$\begin{aligned} U_{v,k} \leftarrow &= U_{v,k} + 2\mu V_{j,k}(M_{v,j} - \Sigma_{k=1}^r U_{v,k}V_{j,k}) \\ V_{v,k} \leftarrow &= V_{v,k} + 2\mu U_{j,k}(M_{v,j} - \Sigma_{k=1}^r U_{v,k}V_{j,k}) \end{aligned}$$

(b) To avoid overfitting, we usually add regularization terms, which penalize for large values in U and V. Redo part (a) using the regularized objective function below.

$$\frac{\partial E(U,V)}{\partial U_{v,k}} = -2V_{j,k}(\partial(M_{v,j} - \Sigma_{k=1}^r U_{v,k}V_{j,k}) + 2\lambda U_{v,k}$$

$$\frac{\partial E(U,V)}{\partial V_{v,k}} = -2U_{j,k}(\partial(M_{v,j} - \Sigma_{k=1}^{r} U_{v,k}V_{j,k}) + 2\lambda V_{v,k}$$

thus, formula in (6) can be updated to :

$$U_{v,k} \leftarrow = U_{v,k} + 2\mu V_{j,k}(M_{v,j} - \Sigma_{k=1}^{r} U_{v,k}V_{j,k}) - 2\lambda U_{v,k}$$
$$V_{v,k} \leftarrow = V_{v,k} + 2\mu U_{j,k}(M_{v,j} - \Sigma_{k=1}^{r} U_{v,k}V_{j,k}) - 2\lambda V_{v,k}$$

(c) Implement *myRecommender.m* by filling the gradient descent part.
Report:
The following table shows the RMSE of $myRecommender3.m$ file.

lowRank	Training RMSE	Testing RMSE	logTime
3	0.845	0.9319	42.56
5	0.7962	0.9499	41.98
7	0.7598	0.9553	42.87
9	0.7243	0.9788	46.95

Observations:
The RMSE of training set is decreasing when the number of lowRank is increasing, however, RMSE of testing set does not decrease.