Longxiang Wang
12710172
wlongxiang1119@gmail.com

Homework Assignment 1
Machine Learning 1, 19/20

2019-09-13

1 Multivariate Calculus

1.1 First Subtask

a)

Suppose we have

$$\boldsymbol{y = x - \mu}$$

Then we can simplify the original form to:

$$\boldsymbol{\nabla_\mu y^T \Sigma^{-1} y}$$

According to chain rule:

$$\boldsymbol{\nabla_\mu y^T \Sigma^{-1} y = (\nabla_y y^T \Sigma^{-1} y) \nabla_\mu y}$$

And we can derive the first part of the above form by:

$$\begin{aligned} &\boldsymbol{\nabla_y y^T \Sigma^{-1} y} \\ &\boldsymbol{= y^T \nabla_y (\Sigma^{-1} y) + \nabla_y y^T (\Sigma^{-1} y)} \\ &\boldsymbol{= 2 y^T \Sigma^{-1}} \end{aligned}$$

by plugging in y and $\nabla_\mu y$ we have the final solution:

$$\boldsymbol{\nabla_\mu y^T \Sigma^{-1} y = -2(x - \mu)^T \Sigma^{-1}}$$

b)

It is easy to see that the function is real-valued function. So, $\frac{\delta f}{\delta \boldsymbol{q}} \in \mathbb{R}^{1 \times n}$

$$[\frac{\delta f}{\delta q_1}, \frac{\delta f}{\delta q_2}, ..., \frac{\delta f}{\delta q_j}, \frac{\delta f}{\delta q_n}] = [\frac{p_1}{q_1}, \frac{p_2}{q_2}, ..., \frac{p_j}{q_j}, \frac{p_n}{q_n}]$$

c)

We have

$$\boldsymbol{f = Wx}, \quad f \in \mathbb{R}^2, \quad \boldsymbol{W} \in \mathbb{R}^{2 \times 3}, \quad \boldsymbol{x} \in \mathbb{R}^3$$

So the derivative is of shape:

$$\frac{\boldsymbol{df}}{\boldsymbol{dw}} \in \mathbb{R}^{2 \times (2 \times 3)}$$

Then, we have:

$$\frac{\boldsymbol{df}}{\boldsymbol{dw}} = \begin{bmatrix} \frac{\boldsymbol{df_1}}{\boldsymbol{dw}} \\ \frac{\boldsymbol{df_2}}{\boldsymbol{dw}} \end{bmatrix}, \quad \frac{\boldsymbol{df_i}}{\boldsymbol{dw}} \in \mathbb{R}^{1 \times (2 \times 3)}$$

If, we expand the matrices and apply elementwise calculation, we should have the partial derivative:

$$\frac{\delta f_i}{\delta W_{iq}} = x_q$$

This allows us to compute the partial derivatives of f_i with respect to a row of W, which is given as

$$\frac{\delta f_i}{\delta W_{i,(1,2,3)}} = \boldsymbol{x^T} \in \mathbb{R}^{1\times(1\times3)}$$

And for other rows:

$$\frac{\delta f_i}{\delta W_{k\neq i,(1,2,3)}} = \boldsymbol{0^T} \in \mathbb{R}^{1\times(1\times3)}$$

So finally, we have:

$$\frac{\delta f_1}{\delta W} = \begin{bmatrix} x_1 & x_2 & x_3 \\ 0 & 0 & 0 \end{bmatrix}$$
$$\frac{\delta f_2}{\delta W} = \begin{bmatrix} 0 & 0 & 0 \\ x_1 & x_2 & x_3 \end{bmatrix}$$

d)

We have:

$$W \in \mathbb{R}^{m\times k}, \quad x \in \mathbb{R}^{k\times 1}, \quad f \in \mathbb{R}, \quad \nabla_W f \in \mathbb{R}^{m\times k}$$

We can use the solutions derived above to calculate:

$$\boldsymbol{\nabla_W f = 2(\mu - Wx)^T \Sigma^{-1} \nabla_W (Wx)}$$

Where we have $\boldsymbol{\nabla_W (Wx)}$:

$$\boldsymbol{\nabla_W (Wx) = \begin{bmatrix} \frac{df_1}{dw} \\ \frac{df_2}{dw} \\ \dots \\ \frac{df_m}{dw} \end{bmatrix}, \quad \frac{df_i}{dw} \in \mathbb{R}^{1\times(m\times k)}}$$

Where

$$\boldsymbol{\frac{df_i}{dW} = \begin{bmatrix} 0^T \\ \dots \\ 0^T \\ x^T \\ 0^T \\ \dots \\ 0^T \end{bmatrix}}$$

2 Probability Theory

2.1 First Subtask

a)

Even though criminal rate in population is relatively low (the prior), based on the observation, the chance of seeing a criminal under such situation (posterior) is very high.

b)

We have:

$$P(criminal) = \frac{1}{10^5}$$

$$P(observedSituation|criminal) = 0.8$$

$$P(observedSituation|notCriminal) = \frac{1}{10^6}$$

With this, according to the sum rule, we can calculate the probability of observing such situation:

$$\begin{aligned} P(observedSituation) = P(observedSituation|criminal)P(criminal)+ \\ P(observedSituation|notCriminal)P(notCriminal) \\ = \frac{0.8}{10^5} + \frac{1 - \frac{1}{10^5}}{10^6} \end{aligned}$$

According to Bayes:

$$P(criminal|observedSituation) = \frac{P(observedSituation|criminal)P(criminal)}{P(observedSituation)} = 0.89$$

c)

See solution is b).

d)

With the new evidence, the probability of making this observation becomes much higher, thus according to Bayes rule as shown in b), the belief in the man in question being a criminal lowers down. the probability of making this observation when the man is not a criminal is 1

2.2 Second Subtask

a)

Assume $\boldsymbol{\rho}$ satisfies i.i.d., and each oberservation also satisfies i.i.d., we have:

$$P(\boldsymbol{D}|\boldsymbol{\rho}) = \prod_{n=1}^{N} \prod_{i=1}^{4} (\boldsymbol{\rho})^{\boldsymbol{x_{ni}}}$$

b)

Since we have 4 out of 8 ♡ and also 4 out of 8 ♠, and none of other suits in the observations. The max likelihood solution is as follows:

$$\boldsymbol{\rho} = [1/2, 0, 0, 1/2]$$

c)

Suppose we have M red observations in N:

$$P_{ML}(D|p) = p^M (1-p)^{N-M}$$

d)

Since we have 6 out of 8 in red, so the solution that maximize likelihood is $3/4$.

e)

$$p = \frac{\rho_1 + \rho_3}{\rho_1 + \rho_2 + \rho_3 + \rho_4}$$

f)

$$P_{posterior}(p_{bernoulli}|D) = \frac{p_{likelihood}(D|p_{bernoulli})p_{bernoulli}}{p_{evidence}(D)}$$

Where $p_{bernoulli}$ is the prior, posterior, likelihood and evidence are pointed out in the form.

g)

$$p_{map} = argmax_p P(D|p)P(p) = argmax_p p^M (1-p)^{N-M} p^{\alpha-1}(1-p)^{\beta-1}$$

We find p such that the derivative of posterior:

$$\frac{d(p^M (1-p)^{N-M} p^{\alpha-1}(1-p)^{\beta-1})}{dp} = -p(N-M+\beta-1) + (M+\alpha-1)(1-p) = 0$$

Finally we have:

$$p = \frac{1-\alpha-M}{2-N-\beta-\alpha}$$