SUSTech

MACHINE-LEARNING

term

Homework 2

刘禹熙 2020 年 10 月 26 日

2.1. (a) True or False If two sets of variables are jointly Gaussian, then the conditional distribution of one set conditioned on the other is again Gaussian. Similarly, the marginal distribution of either set is also Gaussian.

Solution. True

(b) We consider a partitioning of the components of x into three groups x_a, x_b, and x_c, with a corresponding partitioning of the mean vector μ and of the covariance matrix $\sum$ in the form

$$\mu = \begin{bmatrix} \mu_a \\ \mu_b \\ \mu_c \end{bmatrix}, \Sigma = \begin{bmatrix} \Sigma_{aa} & \Sigma_{ab} & \Sigma_{ac} \\ \Sigma_{ba} & \Sigma_{bb} & \Sigma_{bc} \\ \Sigma_{ca} & \Sigma_{cb} & \Sigma_{cc} \end{bmatrix}$$

Find an expression for hte conditional distribution $p(x_a|x_b)$ in which x_c has been marginalized out.

Solution.

$$p(x_a, x_b) = \int p(x_a, x_b, x_c) dx_c = N(x|\mu, \Sigma)$$

$$\mu = \begin{bmatrix} x_a \\ x_b \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{bmatrix}$$

$$\Lambda = \Sigma^{-1} = \begin{bmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{bmatrix}$$

So the $p(x_a|x_B)$ is

$$p(x) = \mathcal{N}(x|\mu, \Sigma)$$

$$x = \begin{bmatrix} x_a \\ x_b \end{bmatrix} \mu = \begin{bmatrix} \mu_a \\ \mu_b \end{bmatrix} \Sigma = \begin{bmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{bmatrix}$$

$$\mu_{a|b} = \mu_a + \Sigma_{ab}\Sigma_{bb}(X_B - \mu_b)$$

$$\Sigma_{a|b} = \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba}$$

2.2. Consider a joint distribution over the variable

$$z = \begin{bmatrix} x \\ y \end{bmatrix}$$

whose mean and covariance are given by

$$\mathbb{E}[z] = \begin{bmatrix} \mu \\ A\mu + b \end{bmatrix}, cov[z] = \begin{bmatrix} \Lambda^{-1} & \Lambda^T \\ A\Lambda^{-1}L^{-1} & L^{-1} + A\Lambda^{-1}A^T \end{bmatrix}.$$

(a) Show that the marginal distribution p(x) is given by p(x) = $\mathcal{N}(x|\mu, \Lambda^{-1})$.

Solution. First we need to prove that $p(x)$ also obeys Gaussian distribution. And we need to eliminate y. As mentioned before. We assume that

$$\begin{aligned} \Lambda_a a &= \Lambda^{-1} \\ \Lambda_a b &= \Lambda^T \\ \Lambda_b a &= A\Lambda^{-1} \\ \Lambda_{bb} &= L^{-1} + A\Lambda^{-1}A^T \end{aligned}$$

$$\begin{aligned} -\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu) = &-\frac{1}{2}(x_a-\mu_a)^T\Lambda_{aa}(x_a-\mu_a) - \frac{1}{2}(x_a-\mu_a)^T\Lambda_{ab}(x_b-\mu_b) \\ &-\frac{1}{2}(x_b-\mu_b)^T\Lambda_{ba}(x_a-\mu_a) - \frac{1}{2}(x_b-\mu_b)^T\Lambda_{bb}(x_b-\mu_b) \end{aligned}$$

Take out item that only involve x_b

$$-\frac{1}{2}x_b^T\Lambda_{BB}x_b + x_b^T m = -\frac{1}{2}(x_b - \Lambda_{aa}^{-1}m)^T\Lambda_{bb}(x_b - \Lambda_{bb}^{-1}) + \frac{1}{2}m^T\Lambda_{bb}^{-1}m$$

$$m = \Lambda_{bb}\mu_b - \Lambda_{ba}(x_a - \mu_a)$$

take the quadratic term substituted into the above formula

$$\int \exp\{-\frac{1}{2}(x_b - \Lambda_{aa}^{-1}m)^T\Lambda_{bb}(x_b - \Lambda_{bb}^{-1})\}dx_b$$

This is the inverse of a Gaussian distribution and has no effect on the result. The rest of the equation term combined with the above formula is :

$$\begin{aligned} &\frac{1}{2}[\Lambda_{bb}\mu_b - \Lambda_{ba}(x_a-\mu_a)]^T\Lambda_{bb}^{-1}[\Lambda_{bb}\mu_b - \Lambda_{ba}(x_a-\mu_a)] \\ &-\frac{1}{2}x_a\Lambda_{aa}x_a + x_a^T(\Lambda_{aa}\mu_a + \Lambda_{ab}\mu_b) + C \\ = -\frac{1}{2}x_a^T&(\Lambda_{aa} - \Lambda_{ab}\Lambda_{bb}^{-1}\Lambda_{ba})^{-1} + x_a^T(\Lambda_{aa} - \Lambda_{ab}\Lambda_{bb}^{-1}\Lambda_{ba})^{-1}\mu_a + C \end{aligned}$$

Compare with the above formula

$$\Sigma_a = (\Lambda_{aa} - \Lambda_{ba}\Lambda_{bb}^{-1}\Lambda_{ba})^{-1}$$

We have:

$$\begin{aligned} \mathbb{E}[x_a] &= \mu_a \\ cov[x_a] = \Sigma_{aa} = \Lambda_{aa} &= \Lambda^{-1} \\ p(x) &= \mathcal{N}(x|\mu, \Lambda^{-1}) \end{aligned}$$

(b) Show that the conditional distribution $p(y|x)$ is given by $p(y|x) = \mathcal{N}(y|Ax + b, L^{-1})$.

Solution. 内容...

2.3. Show that the convariance matrix Σ that maximizes the log likelihood function is given by the sample covariance

$$\Sigma_{ML} = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu_{ML})(x_n - \mu_{ML})^T$$

Is the final result symmetric and positive definite(provided the sample covariance is nonsingular)?

Solution. We have

$$p(X|\mu, \Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} e^{-(x-\mu)^T\Sigma^{-1}(x-\mu)/2}$$

Given i.i.d. data $X = (x_1, ..., x_n)^T$, the log likelihood function is given by

$$\ln p(X|\mu, \Sigma) = -\frac{ND}{2}\ln(2\pi) - \frac{N}{2}\ln|\Sigma| - \frac{1}{2}\sum_{n=1}^{N}(x_n - \mu)^T\Sigma^{-1}(x_n - \mu)$$

Set the derivative of the log likelihood function to zero,

$$\frac{\partial}{\partial\mu}\ln p(X|\mu, \Sigma) = \sum_{n=1}^{N}\Sigma^{-1}(x_n - \mu) = 0$$

$$\frac{\partial}{\partial\Sigma}\ln p(X|\mu, \Sigma) = \frac{N}{2}(\Sigma^{-1} - \Sigma^{-1}\frac{1}{N}\sum_{n=1}^{N}(x^n - \mu_{\hat{ML}})(x^n - \mu_{ML})^T\Sigma^{-1})^T = 0$$

solve to obtain

$$\hat{\mu_{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

$$\hat{\Sigma_{ML}} = \frac{1}{N}\sum_{n=1}^{N}(x_n - \hat{\mu_{ML}})(x_n - \hat{\mu_{ML}})^T$$

2.4. (a) Derive an expression for the sequential estimation of the variance of a univariate Gaussian distribution, by starting with the maximum likelihood expression

$$\sigma^2_{ML} = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu)^2$$

Verify that substituting the expression for a Gaussian distribution into the Robbins-Monro sequential estimation formula gives a result of the same form, and hence obtain an expression for the corresponding coefficients a_N.

Solution.

$$L(x_1, x_2, ...x_n|\mu, \sigma^2) = \prod_{i=1}^{N}\frac{1}{\sqrt{2\pi}\sigma}\exp^{\frac{(x_i-\mu)^2}{2\sigma^2}}$$

$$\ln L(x_1, x_2, ...x_n|\mu, \sigma^2) = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{2}\ln\sigma^2 - \frac{N}{2}\ln(2\pi)$$

$$\frac{\partial \ln L}{\partial \sigma^2} = \frac{1}{2\sigma^4}\sum_{n=1}^{N}(x_n - \mu)^2 - \frac{N}{2\sigma^2} = 0$$

$$\sigma^2_{ML} = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu)^2$$

Robbins-Monro:

$$\frac{\partial}{\partial\sigma^2}\frac{1}{N}\sum_{n=1}^{N} -\ln L(x_n|\sigma^2) = 0$$

$$\lim_{n\to\infty}\frac{1}{N}\sum_{n=1}^{N}\frac{\partial}{\partial\sigma^2}\ln L(x_n|\sigma^2) = E_x[-\frac{\partial}{\partial\sigma^2}\ln L(x|\sigma^2)]$$

On the right side of the equation, we use Robbins-Monro

$$\sigma^2_{(N)} = \sigma^2_{(N-1)} - \alpha_{N-1}\frac{\partial}{\partial\sigma^2_{(N-1)}}[-\ln L(x_n|\sigma^2)]$$

$$z = -\frac{\partial}{\partial \sigma_{ML}^2} \ln L(x|\mu_{ML}, \sigma^2) = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu_{ML})^2$$

(b) Derive an expression for the sequential estimation of the covariance of a multivariate Gaussian distribution, by starting with the maximum likelihood expression

$$\Sigma_{ML} = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu_{ML})(x_n - \mu_{ML})^T.$$

Verify that substituting the expression for a Gaussian distribution into the Robbins-Monro sequential estimation formula gives a result of the same form, and hence obtain an expression for the corresponding coefficients a_N.

2.5. Consider a D-dimensional Gaussian random variable x with distribution $N(x|\mu, \Sigma)$ in which the covariance Σ is known and for which we wish to infer the mean μ from a set of observations $X = \{x_1, x_2, ..., x_N\}$. Given a prior distribution $p(\mu) = N(\mu|\mu_0, \Sigma_0)$, find the corresponding posterior distribution $p(\mu|X)$

Solution.

$$p(\mu|X) \propto p(X|\mu)p(\mu)$$

we can get posterior distribution

$$p(\mu|X) = \mathcal{N}(\mu|\mu_N, \Sigma_N)$$

$$\mu_N = \frac{\Sigma}{N\Sigma_0 + \Sigma}\mu_0 + \frac{N\Sigma_0}{N\Sigma_0 + \Sigma}\mu_{ML}$$

$$\frac{1}{\Sigma_N} = \frac{1}{\Sigma_0} + \frac{N}{\Sigma}$$

$$\hat{\mu_{ML}} = \frac{1}{N}\sum_{n=1}^{N} X_n$$

2.6. program question

(a) How could having a larger dataset influence the performance of KNN?

Solution. The execution time of the algorithm will greatly increase, beacuse the KNN algorithm needs to select the nearest K points, it first needs to calcution the distance among the all points, and then sort them by distance. The time complexity is $O(nlogn)$. As the amount of the data increase, the execution time will increase significantly.

(b) Tabulate your results in Table 1 for the validation set as shown below and include that in your file.

K	Norm	Accuarcy(%)
3	L1	88.4
3	L2	88.4
3	L-inf	89.8
5	L1	91.3
5	L2	89.8
5	L-inf	89.8
7	L1	89.8
7	L2	91.3
7	L-inf	89.8

(c) Finally, mention the best K and the norm combination you have settled upon from the above table and report the accuracy on the test set using that combination.

Solution. In the above table, we can see when the K=5& L1 norm and k=7 &l2 norm we can get the best accuracy with 91.3%

Solution. $L(w) = \prod_{i=1}^{n} p(Y^i|X^i;w)^{Y^i}(1-h(X^i;w))^{1-Y^i}$

$l(w) = \ln L(w) = \sum_{i=1}^{n} Y^i \ln(h(X^i;w)) + (1-Y^i)\ln(1-h(X^i;w))$

$\frac{\partial}{\partial w_j} l(w) = (y - h(X;w))x_j$