HW1 Report

309555025 羅文笙

1 Bayesian Linear Regression

1-1. Why we need basis function? What is the benefit for applying it over linear regression?

Ans: Sometimes, only using the input data is not enough to correctly model the linear regression problem. The benefit of it is that by using different basis function, we can adjust the equation adapting to specific linear regression problem, even though the problem is a nonlinearly separable one.

1-2.

1-2 Prove that $p(t|x, \mathbf{x}, \mathbf{t}) = N(t|m(x), s^2(x))$

pf:

1° $p(w|\mathbf{x}, \mathbf{t}) \propto p(\mathbf{t}|\mathbf{X}, w)\, p(w|\alpha)$

by P.93 的公式

$p(t|\mathbf{X}, w) = N(t|w^T \phi(x), \beta^{-1} I) = N(t|w^T A + b, L^{-1})$

$\Rightarrow A = \phi(x)^T,\ b = 0,\ L = \beta I$ ..... (1)

$p(w|\alpha) = N(w|0, \alpha^{-1} I) = N(w|m, \Lambda^{-1})$

$\Rightarrow m = 0,\ \Lambda = \alpha I$ ..... (2)

$p(w|x,t) = N(w|\Sigma(A^T L(w-b) + \Lambda m), \Sigma)$, where $\Sigma = (\alpha I + A^T L A)^{-1}$

使用 (1), (2) 去代換，得

$N(w|S(\phi(x)^T \beta t), S)$

2° $p(t|w, x) = N(t|w^T \phi(x), \beta^{-1}) = N(t|w^T A + b, L^{-1})$

$\Rightarrow A = \phi(x),\ b = 0,\ L = \beta I$ ..... (3)

$p(w|x,t) = N(w|S(\beta\phi(x)t), S) = p(w|m, \Lambda^{-1})$

$\Rightarrow m = S(\beta\phi(x)t),\ \Lambda^{-1} = S$ ..... (4)

使用 (3), (4) 去代換，得

$p(t|x, \mathbf{x}, \mathbf{t}) = N(t|Am + b, L^{-1} + A\Lambda^{-1}A^T)$

$= N(t|\beta\phi(x)^T S\phi(x)t, \beta^{-1} + \phi(x)^T S\phi(x))$

(By 題目給的 $m(x)$, $s^2(x)$ 定義)

$= N(t|m(x), s^2(x))$ 共 得証

1-3. Could we use linear regression function for classification? Why or why not? Explain it!

Ans: No, we could not use linear regression for classification. First, we assume the target we want to predict is a binary label data denote by 0 and 1, where 0 denote “No” and 1 denote “Yes”. If predicted value is greater than 0.5, predict “Yes”, otherwise, predict “No”. There are mainly two problems for applying linear regression on classification problem.

Problem1: By applying linear regression method on this problem with known dataset, we will get a straight line which might range outside 0 and 1. Since the

problem we want to solve is probabilistic not continuous, output greater than 1 or lower than 0 doesn't really make sense.

Problem2: Given that he label of our training data is either 0 or 1, if the training data is imbalance, for example, f(1) = 0, f(10000) = 0, the best fitting line of linear regression model will be influence by those imbalance data and cannot correctly predict some inputs.

2 Linear Regression(By using Normal equation)

2-1(a)

	M = 1	**M = 2**
RMS error of training set	0.063	0.0489
RMS error of valid set	0.0601	0.0489

2-1(b)

	Weight in M=1	**Square value of weight**
GRE score	0.036	0.0013
TOFEL score	0.017	0.0003
University rating	0.008	0.000079
SOP	0.0003	0.000000123
LOR	0.0227	0.0005
CGPA	0.0618	**0.0038**
Research	-0.017	0.0003

Explanation:

If we use weights f polynomial model M = 1 to select the most contributive feature, it seems that 'CGPA' will be a good choice, since its square value of weight is the greatest, which means that 'CGPA' influences the predicted value of model the most.

2-2(a)

Given that the 7 features of input data might have some dependency between them, I would use polynomial basis function to further improve my regression model.

2-2(b)

	M = 1	Polynomial basis function
RMS error of training set	0.0631	0.0601
RMS error of valid set	0.0429	0.0489

As the experimental result shown, RMS error of training set decreases from '0.0631' to '0.0601', which means our new model has learned more dependency in training data. However, RMS error of valid set increases from '0.0429' to '0.0489', which means the new model becomes too complex and over-fits the training data.

2-2(c)

	RMS error of training set	**RMS error of valid set**
M = 1	0.0587	0.0594
M = 2	0.0557	0.0639

By applying k-fold cross-validation(with k = 5), I select order M = 1. Although the model of order M = 2 has a better RMS error of training set, its RMS error of valid set is bigger than the model of order M = 1. The reason is that it might be due to the overfitting problem caused by too many parameters in model of order M = 2.

2-3(a)

Maximum likelihood approach is good for linear regression. However, if we have enough parameters, it will generate parameters with high magnitude which will span a function oscillating widely and passing through each data points. In consequence of that, we cannot achieve a good generalization with maximum likelihood approach, so I think 'generalization' is the key difference between MLE and MAP

2-3(b)

Polynomial basis function

	RMS error of train set	RMS error of valid set
MLE	0.0601	0.0489
MAP(λ = 1)	0.0604	0.0462
MAP(λ = 100)	0.0616	0.0455
MAP(λ = 10000)	0.0633	0.0434

K-fold(M=1)

	RMS error of train set	RMS error of valid set
MLE	0.0557	0.0639
MAP(λ = 1)	0.0642	0.0652
MAP(λ = 100)	0.0739	0.0760
MAP(λ = 10000)	0.1081	0.1080

2-3(c)

As the experimental results shown in 2-3(b), the RMS errors of valid set of polynomial basis function decrease as we increase the lambda, which means MAP does prevent the overfitting problem in it. However, the RMS errors of valid set of model M= 1(K-fold) didn't show the same result as polynomial basis function. It might because that model M = 1 doesn't have enough parameters to cause serious overfitting problem and the MAP method will make the model become too simple to fit the training set.

So, yes, I believe the experimental results are consistent with my conclusion in 2-3(a).