ML HW1

309513121 李偉齊

1. Bayesian Linear Regression

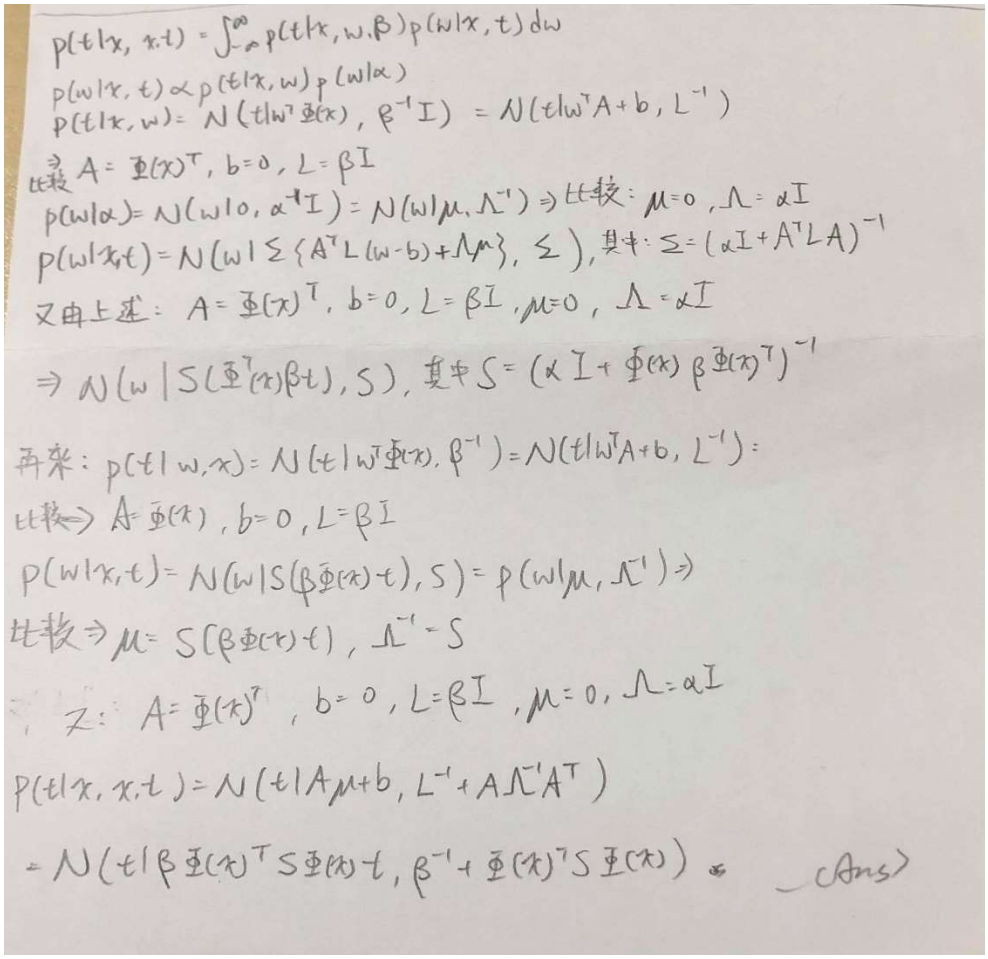
(1) Why we need the basis function ϕ(x) for linear regression? And what is the benefit for applying basis function over linear regression?

Ans :

We use basis function to solve some nonlinear problems. If we use the basis function, the result would be more fit with high dimension model or the basis function we select. We expect to get a better presentation.

(2) Prove that the predictive distribution just mentioned is the same with the form p(t|x, x, t) = N (t|m(x), s^2 (x)) where……

Ans :

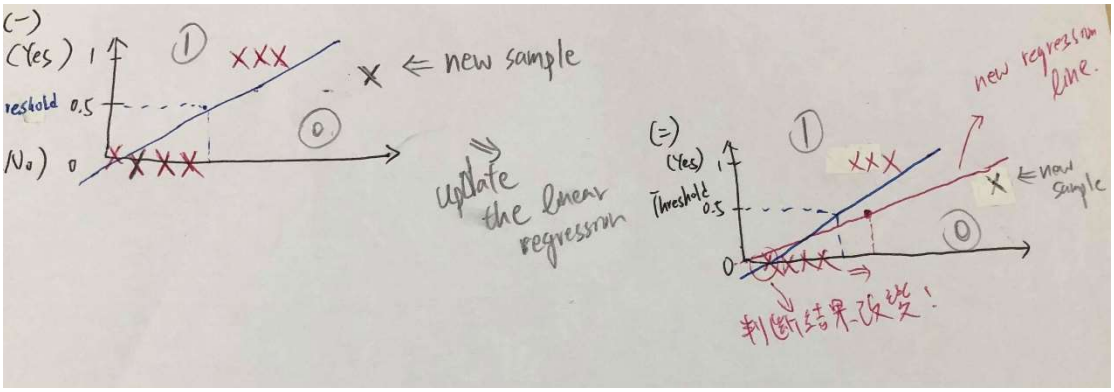


(3) Could we use linear regression function for classification? Why or why not?

Ans :

No, we couldn't use linear regression function for classification. When we use linear regression function, we would get the fitting result. The

classification is determining the result to two or more categories, and the characteristic of line regression ‘fitting’ could cause the wrong determining. As the below picture:



The change of the line would make a mistakes for the classification following the boundary changing.

2. Linear Regression

1. Feature select

(a) In the feature selection stage, please apply polynomials of order M = 1 and M = 2 over the dimension D = 7 input data. Please evaluate the corresponding RMS error on the training set and valid set.

	Training set	Valid set
M=1	0.10550625115617997	0.09642485719796363
M=2	0.2113504061498533	0.19083179971860678

(b) How will you analysis the weights of polynomial model M = 1 and select the most contributive feature?

Ans:

	Weight value
W0 (bias)	-0.3501115624655407
GRE_score	-0.09020892730660802
TOFEL_score	0.17817666672633087
University_rating	-0.12029889355953846
SOP	-0.6398768087817394
LOR	-0.2038588106241749
CGPA	0.21239049118230122
Research	0.1963228109257304

Consider the absolute values of the weights respectively, we know the higher

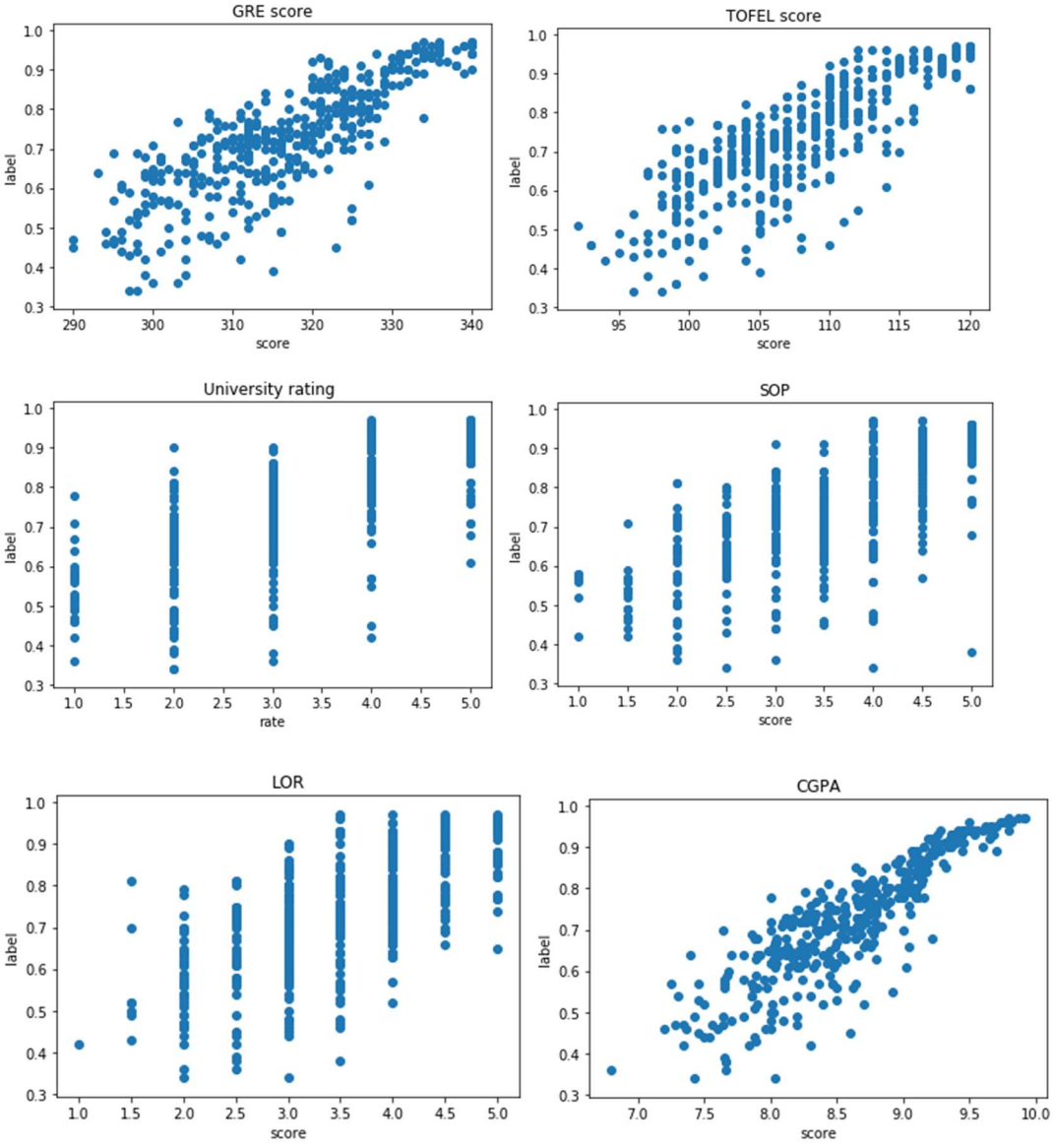
value would affect more in our prediction.
As the result, I select the weight of ‘SOP’: w4 =-0.6398768087817394 to be my answer.

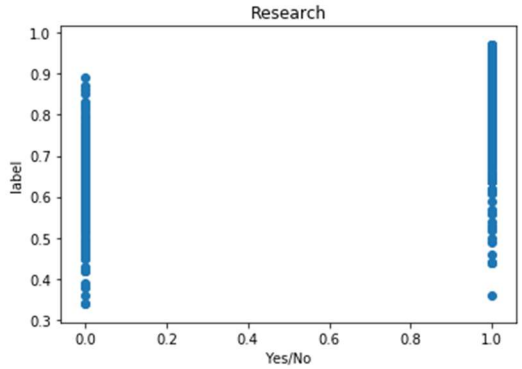
2. Maximum likelihood approach

(a) Which basis function will you use to further improve your regression model, Polynomial, Gaussian, Sigmoidal, or hybrid?

Ans:

I will choose polynomial, because as the plot of the score data compare their labels below, I observed that most of them presented linear.





(b) Introduce the basis function you just decided in (a) to linear regression model and analyze the result you get. (Hint: You might want to discuss about the phenomenon when model becomes too complex.)

Ans:

I select polynomial, so the basis function can present as:

φ(x)={x1,x2,x3,x4,x5,x6,x7,x1x1,x1x2,x1x3,x1x4,x1x5,x1x6,x1x7,x2x2,x2x3,x2x4,x2x5,x2x6,x2x7,x3x3,x3x4,x3x5,x3x6,x3x7,x4x4,x4x5,x4x6,x4x7,x5x5,x5x6,x5x7,x6x6,x6x7,x7x7}

And when the linear regression M=1:

training rms	0.015835573572921993
testing rms	0.01512598692930239

This result is better than above result (2-(a)) which I didn't use polynomial function as basis function.

(P.S.: The reason that I didn't consider the condition M=2 is: the regression would be too complex that the model would be overfitting.)

(c) Apply N-fold cross-validation in your training stage to select at least one hyperparameter(order, parameter number, ...) for model and do some discussion(underfitting, overfitting).

Ans:

I use M=1 regression:

	Training rms	Testing rms
Test data(No.0~100)	0.06887425618065372	0.07856129500538307
Test data(No.100~200)	0.07262741476945624	0.06584689475278374
Test data(No.200~300)	0.07040001730294528	0.05545511262453132
Test data(No.300~400)	0.07463406723520531	0.05854331720838938
Test data(No.400~500)	0.10550625115617997	0.09642485719796363

Consider the above information, we can choose any interval to be our testing data and we still can get a good result on rms.

(P.S.: The result of No.400~500 interval is the answer I used ‘gradient descent’ in the above question (2-(a), M=1). So the rms are higher than another intervals’ rms)

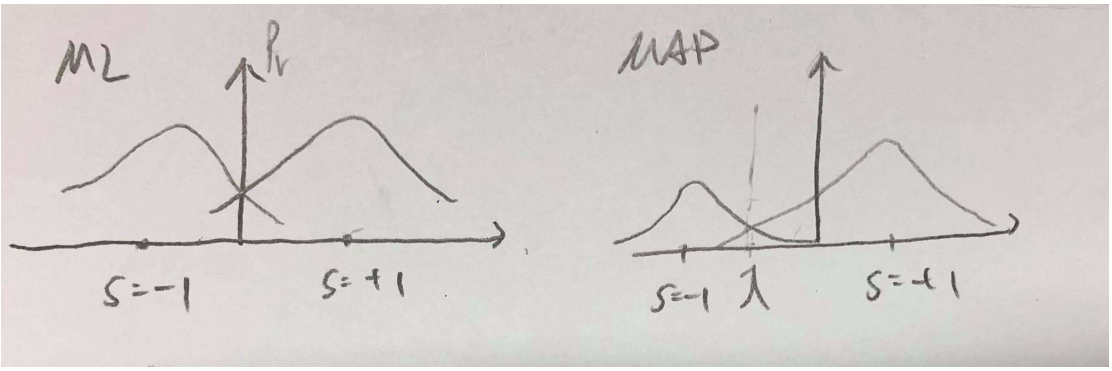
3. Maximum a posterior approach

(a) What is the key difference between maximum likelihood approach and maximum a posterior approach?

Ans:

Assume that we are detecting ‘s’ to be -1 or 1. ML doesn’t consider the situation ‘s’ occur -1 or 1, because ML assume that the probabilities of ‘s’ occur -1 or 1 are the same.

MAP consider the probability that ‘s’ occur -1 or 1. When we use MAP method, we need to decide ‘lambda’ to optimize the detection.



(b) Use Maximum a posterior approach method to retest the model in 2 you designed. You could choose Gaussian distribution as a prior.

Ans:

When we use the ML, we can predict the weights as: w =((A*A)^-1)(A*)b

A matrix is data, A* matrix is the transport of A, b matrix is label.

So, when we use the MAP, we should predict the weights as:

w=((A*A+(lambda)I)^-1)A*b

	Training rms	Testing rms
Lambda = 1	0.07159845722820399	0.06321121534625276
Lambda = 100	0.06100925951684718	0.057437888808838475
Lambda = 10000	0.04515590156661688	0.0434361969888823

(c) Compare the result between maximum likelihood approach and maximum a posterior approach. Is it consistent with your conclusion in (a)?

Ans:
As the result in 3-(b), we find that when the value of lambda larger, the rms of training and testing will be lower. In the conclusion, when we use MAP, the error rate is related to lambda: higher lambda, lower error rate.