Assignment #1

Junpei Zhou 3140102286

1. Machine Learning Problems

(a) Choose words

1) B F

2) C

3) A D

4) C G

5) A E

6) A D

7) B F

8) A E

9) C G

(b)True or False

False. Because this only generates a model which performs well on training dataset, but its performance on the test dataset cannot be guaranteed. In another word, it may leads to "overfitting", and we should split the dataset into training data and validation data to guarantee its ability of generalization.

2. Bayes Decision Rule

(a) Suppose you are given a chance to win bonus grade points

(i) $P(B_1 = 1) = \frac{1}{3}$

(ii) $P(B_2 = 0|B_1 = 1) = 1$

(iii) I think this question is a little ambiguous, if it means "just assume B2 doesn't contain prize, but we are not sure if it contains prize", the probability is: $P(B_1 = 1|B_2 = 0) = \frac{P(B_2=0|B_1=1)*P(B_1=1)}{P(B_2=0)} = \frac{1}{3}/\frac{2}{3} = \frac{1}{2}$, but if it means "the box has already been opened and it doesn't contain the prize", the probability is:
$P(B_1 = 1|B_2 = 0) = \frac{P(B_2=0|B_1=1)*P(B_1=1)}{P(B_2=0)} = \frac{1}{3}/1 = \frac{1}{3}$

(iv) Yes, I should change my choice. Because the probability of the box opened after my choice should also be considered. We can image a scene where I choose box 1 and then box 3 open, and this situation can help us analyze. Suppose that the action I choose box 1 is denoted as B1, and the prize in box i is denoted as Zi, and the box 3 open denoted as O3. So we have:

$$P(O3|Z1, B1) = \frac{1}{2}, P(O3|Z2, 1B) = 1, P(O3|Z3, B1) = 0$$

Then, the conditional probability of getting the prize if I change my choice to box2 is:

$$P(Z2|O3, B1) = \frac{P(O3|Z2,B1)*P(Z2\cap B1)}{P(O3\cap B1)}$$

$$= \frac{P(O3|Z2,B1)*P(Z2\cap B1)}{P(O3|Z1,B1)*P(Z1\cap B1)+P(O3|Z2,B1)*P(Z2\cap B1)+P(O3|Z3,B1)*P(Z3\cap B1)}$$

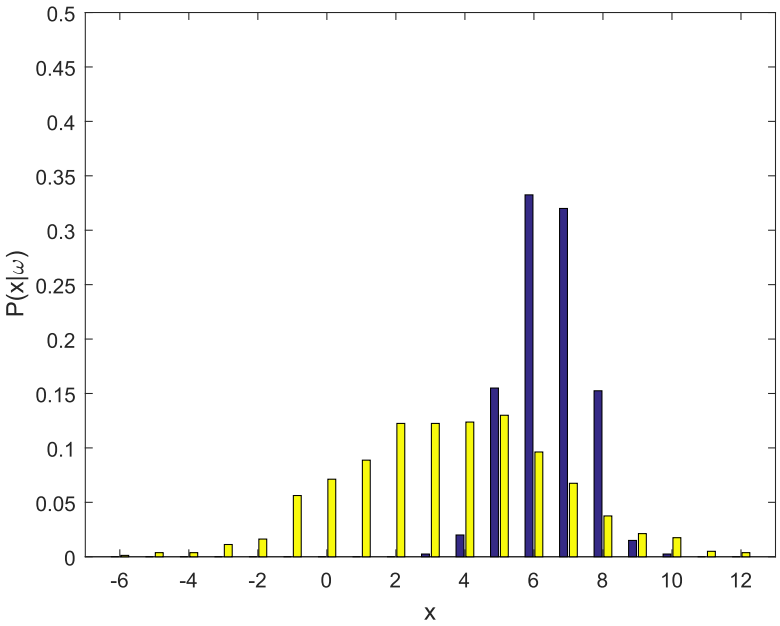
$$= \frac{P(O3|Z2,B1)}{P(O3|Z1,B1)+P(O3|Z2,B1)+P(O3|Z3,B1)} = \frac{1}{1/2+1+0} = \frac{2}{3}$$

Note that the first choice of me is independent of the position of the prize, so they can be eliminated from the numerator and the denominator.

So I will change my choice because its probability of winning the bonus grade points is two times of not changing.

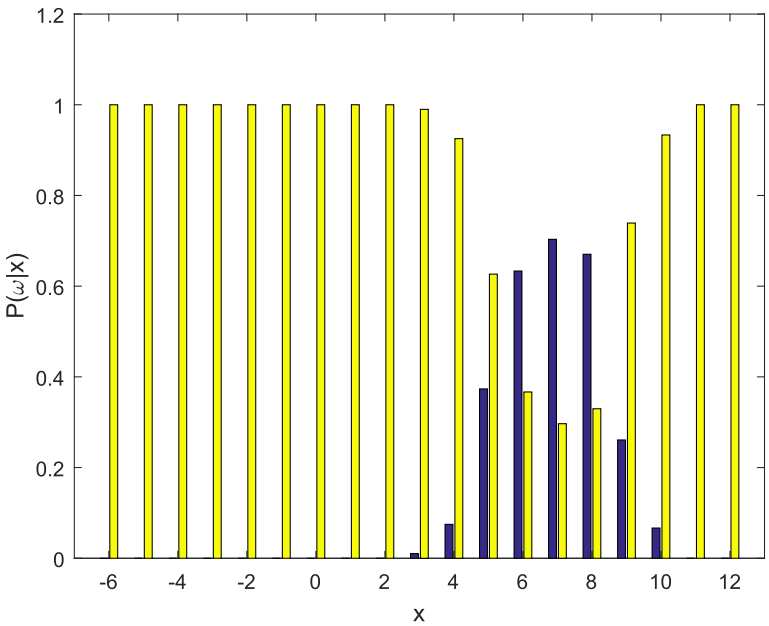
(b) make a two-class classier

(i) The distribution of $P(x|w_i)$ is:



The number of the misclassified sample in test data is 64, and the misclassified rate is 21.33%

(ii) The distribution of $P(w_i|x)$ is:



The number of the misclassified sample in test data is 47, and the misclassified rate is 15.67%

(iii) The minimum total risk is 2.44

$$R(\alpha_1|x) = \lambda_{11}P(w_1|x) + \lambda_{12}P(w_2|x) = P(w_2|x)$$

$$R(\alpha_2|x) = \lambda_{21}P(w_1|x) + \lambda_{22}P(w_2|x) = 2P(w_1|x)$$

$$R = \sum_x min_i R(\alpha_i|x)$$

3. Gaussian Discriminant Analysis and MLE

(a) calculate the posterior probability

$$p(y=1|x;\phi,\mu_0,\mu_1,\Sigma_0,\Sigma_1) = \frac{p(x|y=1;\phi,\mu_0,\mu_1,\Sigma_0,\Sigma_1)*p(y=1)}{p(x|y=0)*p(y=0)+p(x|y=1)*p(y=1)}$$

$$= \frac{N(\mu_1,\Sigma_1)*\phi}{N(\mu_0,\Sigma_0)*(1-\phi)+N(\mu_1,\Sigma_1)*\phi}$$

$$= \frac{1}{e^{-x_1-x_2+1}+1}$$

To find the decision boundary, we denote the discriminant function as

$$g_i(x) = lnP(x|w_i) + lnP(w_i)$$

$$= -\frac{1}{2}(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i) - \frac{d}{2}ln2\pi - \frac{1}{2}ln|\Sigma_i| + lnP(w_i)$$

And the hyperplane defined by the linear equation:

$g_i(x) = g_j(x)$ and here we have $i = 0$ and $j = 1$

Because here we only have two classes, we can get that

$$(x-\mu_0)^T(x-\mu_0) = (x-\mu_1)^T(x-\mu_1)$$

And after solving it, we can get that

$$x_1^2 + x_2^2 = (x_1-1)^2 + (x_2-1)^2$$

So the decision boundary is

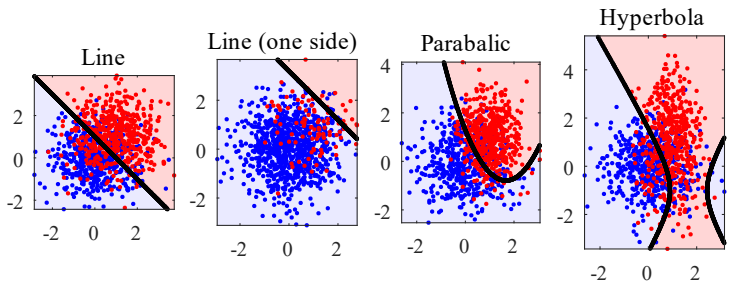
$$x_1 + x_2 = 1$$

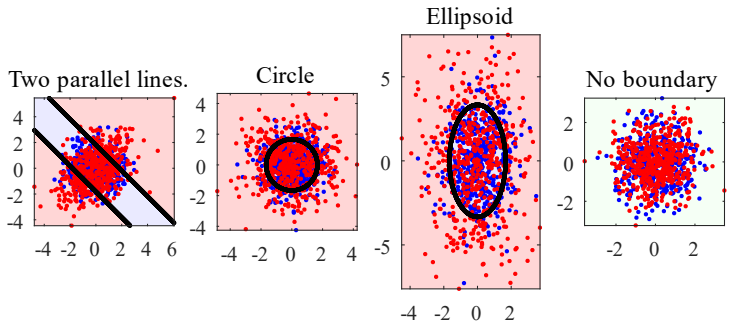
(b) An extension of the above model

The code in *gaussian_pos_prob.m* has been finished. And there is a bug in the template provided, where the return variable is "p", but it declares "P".

(c) Do some field work

Firstly show the result plotted:





The solving process:

Generally speaking, the decision boundary is decided by the equation we got from $g_0(x) = g_1(x)$, and we can get the equation which is shown as below:

$$-\frac{1}{2}(x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0) - \frac{d}{2}ln2\pi - \frac{1}{2}ln|\Sigma_0| + lnP(w_0) =$$

$$-\frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1) - \frac{d}{2}ln2\pi - \frac{1}{2}ln|\Sigma_1| + lnP(w_1)$$

And because we can choose arbitrary Σ_0 and Σ_1, so to be convenient, let us just fix the Σ_0 to be the identical matrix and denote the Σ_1 as [a b; c d]. Then use it into the equation we can get that:

$$(x_1-\mu_{01})^2+(x_2-\mu_{02})^2+ln|\Sigma_1|-2ln\frac{1-\phi}{\phi}=(x-\mu_1)^T\Sigma_1^{-1}(x-\mu 1)$$

Then all the questions below are to choose some a, b, c, d, μ_0, μ_1 and ϕ to make the equation to be some specific shape in the space of x_1 and x_2.

(i) A linear line. We can just use the parameter given in (a), as shown in rum.m.

(ii) A linear line, while both means are on the same side of the line. We can just adjust the ϕ because it determines the intercept of the line. And after solving the inequality, we can get that it must satisfies $e^2<\frac{1-\phi}{\phi}$ or $e^{-2}>\frac{1-\phi}{\phi}$, so we choose $\phi=0.1$

(iii) A parabolic curve. To make the equation to be a parabolic curve, we need to guarantee that the coefficient of x_2^2 is not zero and the coefficient of x_1^2 is zero, and the coefficient of x_1 is not zero, while other coefficient is zero except the coefficient of constant.

(iv) A hyperbola curve. To make the equation to be a hyperbola curve, we need to guarantee that the coefficient of x_1 has the different sign with the coefficient of x_2, liking that if coefficient of x_1 is positive, then the coefficient of x_2 should be negative.

(v) Two parallel lines. As shown in rum.m.

(vi) A circle. To make the equation to be a circle, we need to guarantee that the coefficient of x_1 is equal to the coefficient of x_2

(vii) An ellipsoid. To make the equation to be an ellipsoid, we need to guarantee that the coefficient of x_1 has the same sign with the coefficient of x_2, but their value is not same.

(viii) No boundary, i.e. assigning all samples to only one label. To make no boundary, the distribution of this two kind should be same, as shown in rum.m.

(d)maximum likelihood estimation

$$\phi=\frac{\sum_{i=1}^m y^{(i)}}{m}$$

$$\mu_0=\frac{\sum_{i=1}^m(1-y^{(i)})x^{(i)}}{m-\sum_{i=1}^m y^{(i)}}$$

$$\mu_1=\frac{\sum_{i=1}^m y^{(i)}x^{(i)}}{\sum_{i=1}^m y^{(i)}}$$

4. Text Classication with Naive Bayes

I wrote the code in python and the script is run.py in the /code/text_classification folder.

(a) List the top 10 words that are most indicative of the SPAM class

['nbsp', 'viagra', 'pills', 'cialis', 'voip', 'php', 'meds', 'computron', 'sex', 'ooking']

(b) What is the accuracy of your spam filter on the testing set?

The testing set contains 3011 ham emails and my filter predicts 2986 correctly. The testing set contains 1119 spam emails and my filter predicts 1092 correctly. So my filter predicts correctly on 4078 of 4130 emails.

The overall accuracy of my spam filter on the testing set is 98.74%

(c) True or False, Why?

False.

Because if the ratio of spam and ham email is 1:99, a spam filter can just treat all the emails as ham email and its accuracy is 99%, which is obviously a bad model.

(d) Compute the precision and recall

	SPAM LABEL	HAM LABEL
spam predict	1092	25
ham predict	27	2986

precision = 1092/(1092+25) = 97.76%

recall = 1092/(1092+27) = 97.59%

(e) Which one do you think is more important?

For a spam filter, I think precision is more important. Because if a spam email has not been filtered, it will not bring huge cost to the user to delete it by hand. However, if a very important ham email has been filtered, it can bring a huge cost to the user.

For the classifier to identify drugs and bombs at airport, I think recall is more important. Because if only one bomb that has not been detected by the classifier, the result is catastrophic, which means all the bombs should be detected, even with something else detected as bombs.