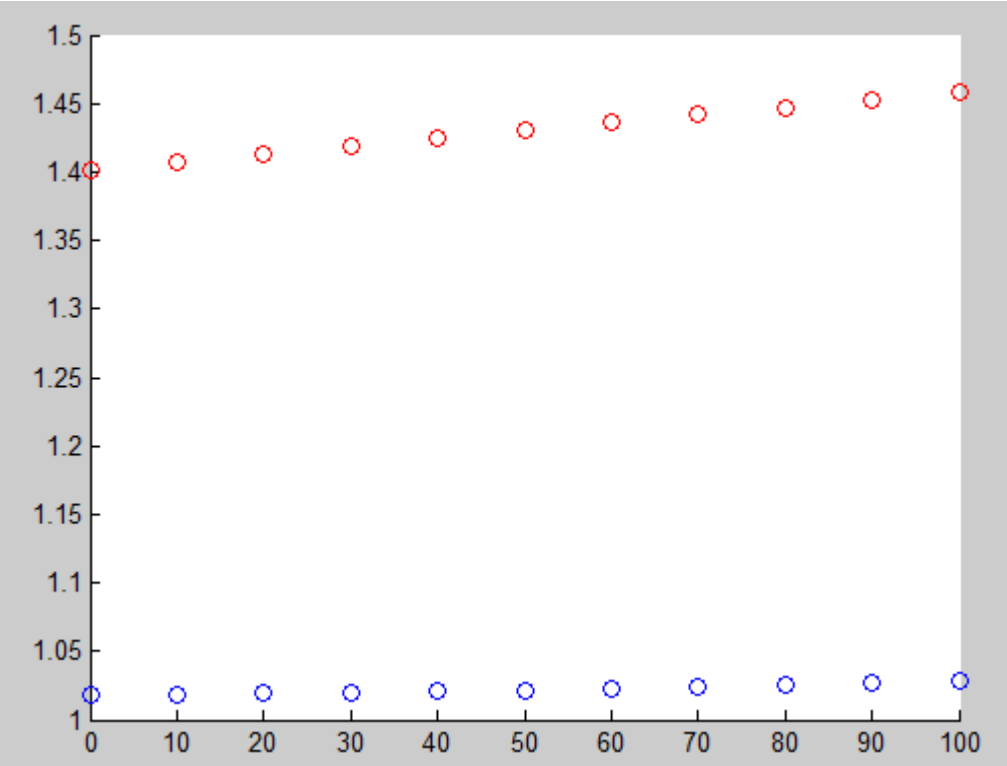
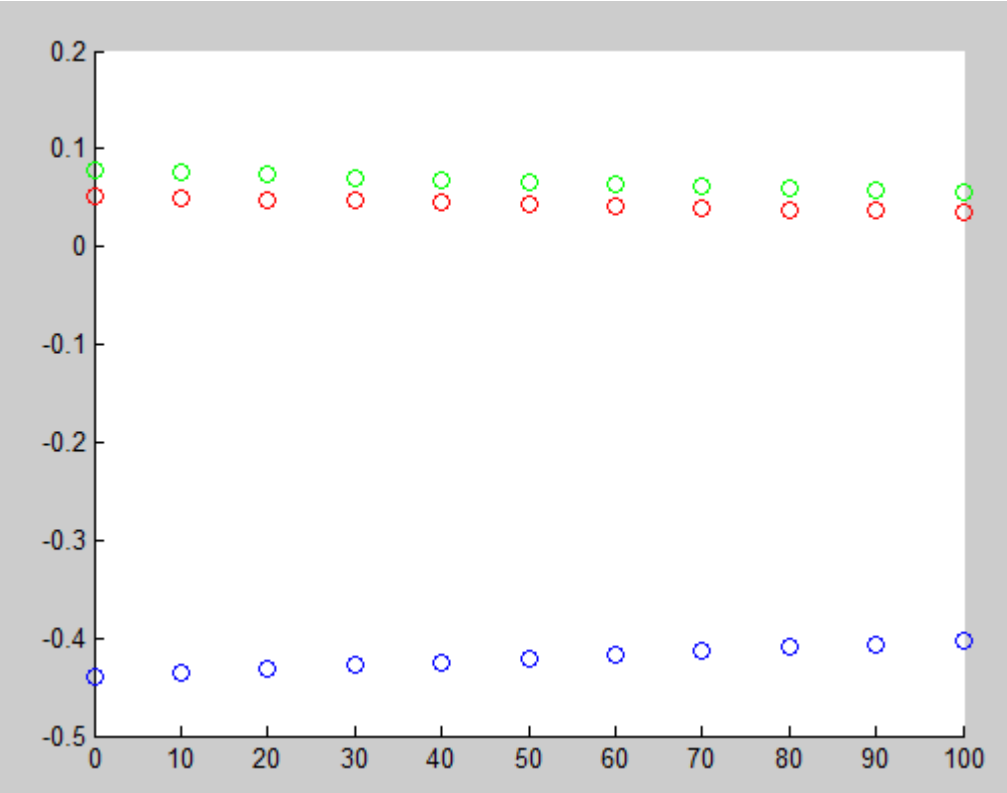
1. [25 points] L1 vs. L2 Regularization
In this problem, we explore the differences between L1 and L2 regularization, discussed in lecture 2. You are given two data _les, hw2x.dat and hw2y.dat (containing inputs and outputs, respectively).
(a) [10 points] Split the data into a training set, containing 90% of the instances, and a test set, containing 10% of the instances (since this is exploratory, we will not do full cross-validation). Write Matlab code to perform L2 regularization. Plot on one graph the root mean squared error on the training set, and the root mean squared error on the test set, as a function of the regularization parameter λ. Vary _ starting at 0, and go high enough that you can see an \interesting" range of behavior. Plot, on a different graph, all of the weights, as a function of λ

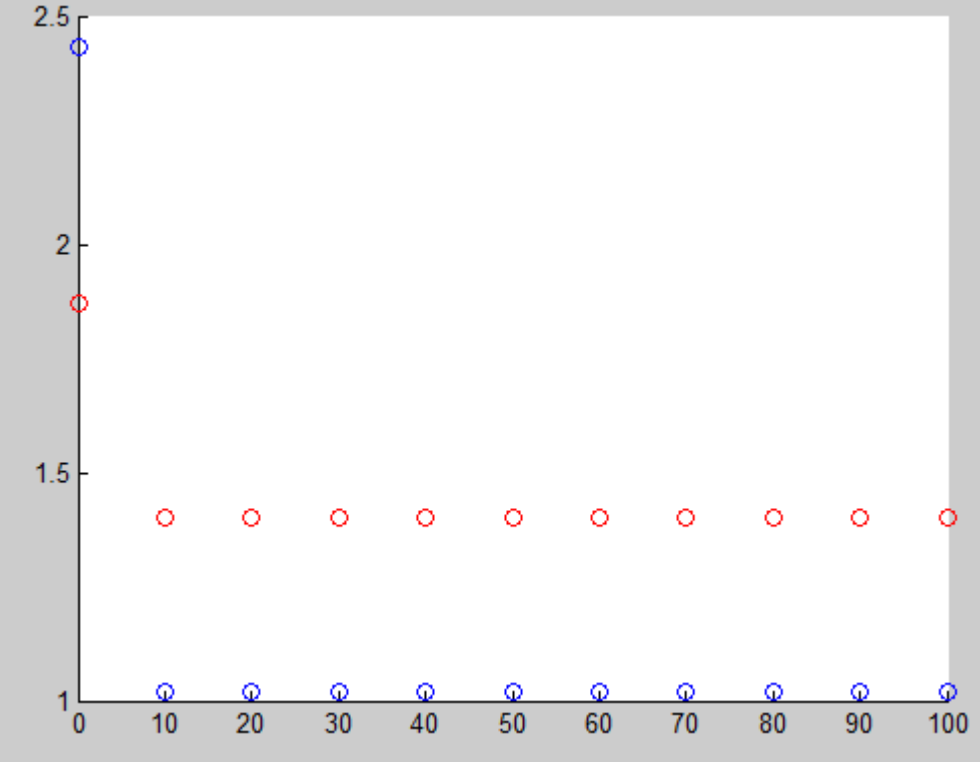


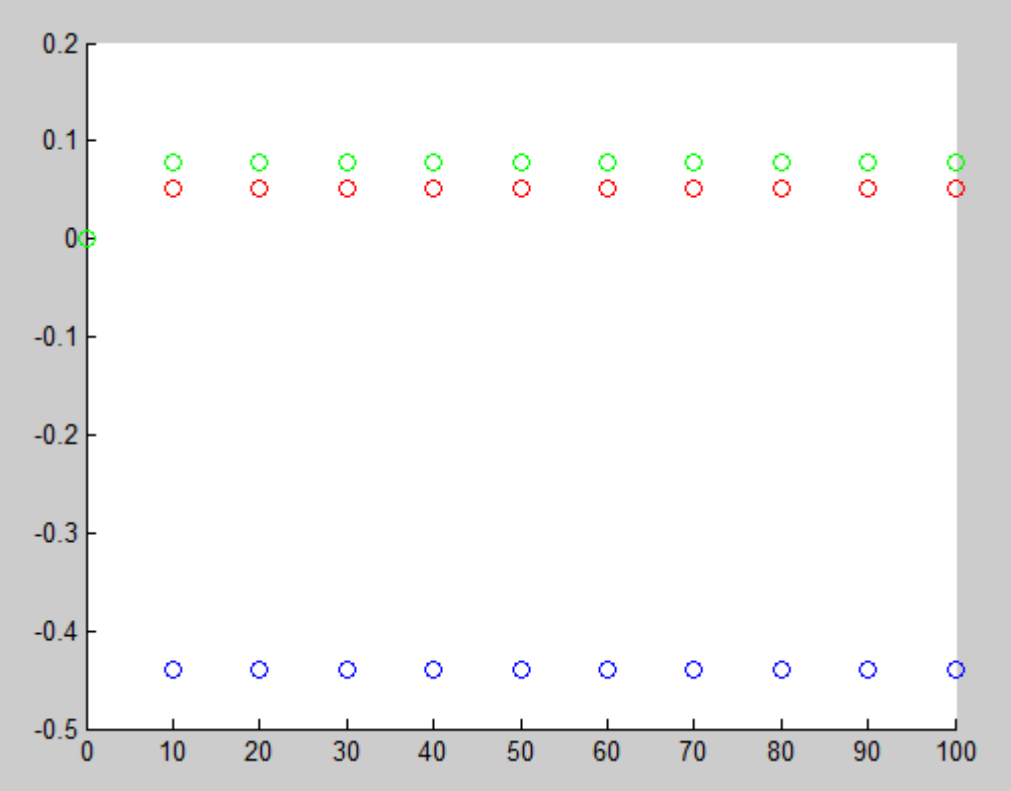


(b) [10 points] Using the quadprog function of Matlab, write a function that performs L1 regularization

see the attached Q1.m

(c) [5 points] Plot the same graphs as above for the L1 regularization. Explain what you observe, and comment on how you think the data was generated.





2. [10 points] Dealing with missing data

Suppose that you use a Gaussian discriminant classifier, in which you model explicitly P(y = 1) (using a binomial) and P(x|y = 0) and P(x|y = 1). The latter have distinct means μ_0 and μ_1, and a shared covariance matrixΣ(a frequent assumption in practice). Suppose that you are asked to classify an example for which you know

inputs $x_1; : : : x_{n-1}$, but the value of x_n is missing. In practice, a common approach in this case is to fill in" the value of x_n by its class-conditional means, $E(x_n|y = 0)$ and $E(x_n|y = 1)$. Using the log-odds ratio, give a mathematical justification for this approach.

$$P(y = 1|x) = \frac{P(x|y = 1)P(y = 1)}{P(x)}$$

Since the latter have distinct means μ_0 and μ_1, and a shared covariance matrix Σ
We have the multivariate Gaussian form for P(x|y = 0) and P(x|y = 1)

$$P(x|y = 0) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x - \mu_0)^T \Sigma^{-1}(x - \mu_0)\right\}$$

$$P(x|y = 1) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x - \mu_1)^T \Sigma^{-1}(x - \mu_1)\right\}$$

3. [10 points] **Naive Bayes assumption**

We discussed in class the fact that naive Bayes assumes that the inputs are conditionally independent given the class label. Suppose now that we have a problem with two binary inputs, x_1 and x_2, which are truly conditionally independent given the class label. Suppose that, by accident, we add a feature, x_3, which is an identical copy of x_2.

(a) [2 points] How many parameters does the initial Naive Bayes classifier have? How many parameters does the classifier with three features have?

a) The parameters of the native Bayes are:

$$\theta_{i,1} = P(x_i|y = 1), \theta_{i,0} = P(x_i|y = 0), \theta_1 = P(y = 1),$$

Hence for two binary input x1, x2, parameters are

$$\theta_1,\ \theta_{1,1}, \theta_{2,1}, \theta_{1,0}, \theta_{2,0}$$

There are 5 parameters.

For two input x1,x2 and duplicate x3, parameters are:

$$\theta_1,\ \theta_{1,1}, \theta_{2,1}, \theta_{1,0}, \theta_{2,0}, \theta_{3,1}, \theta_{3,0},$$

There are 7 parameters.

(b) [8 points] Assuming that the parameters are learned perfectly, describe (using formulas) the effect of the added duplicate feature on the decision boundary of the naive Bayes classifier (by considering the log-odds ratio). What is the worst-case scenario? Explain in 1-2 sentences what this means for the robustness of the naive Bayes classifier.

b)
due to x2, x3 are identical $\theta_{2,1} = \theta_{3,1}$ and $\theta_{2,0} = \theta_{3,0},$

Class label has binomial distribution $P(y) = \theta_1^y(1 - \theta)^{1-y}$ and class conditional

distributions are multivariate Bernoulli

$$P(x|y=1) = \prod_{i=1}^{n} \theta_{i,1}^{xi}(1-\theta_{i,1})^{1-xi}$$

$$P(x|y=0) = \prod_{i=1}^{n} \theta_{i,0}^{xi}(1-\theta_{i,0})^{1-xi}$$

The decision surface:

$$\frac{P(y=1|x)}{P(y=0|x)} = \frac{P(y=1)P(x|y=1)}{P(y=0)P(x|y=0)}$$

$$= \frac{P(y=1)\prod_{i=1}^{n} \theta_{i,1}^{xi}(1-\theta_{i,1})^{1-xi}}{P(y=0)\prod_{i=1}^{n} \theta_{i,0}^{xi}(1-\theta_{i,0})^{1-xi}}$$

Use the lock trick, we get:

$$\log\left(\frac{P(y=1|x)}{P(y=0|x)}\right) = \log\frac{P(y=1|x)}{P(y=0|x)} + \log\frac{P(x_1|y=1)}{P(x_1|y=0)} + \log\frac{P(x_2|y=1)}{P(x_2|y=0)}$$

$$+ \log\frac{P(x_3|y=1)}{P(x_3|y=0)}$$

$\log\frac{P(x_3|y=1)}{P(x_3|y=0)}$ is the additional term added to decision surface, if when the distribution of feature is highly unbalanced, the additional term will impact on the decision surface. For example, the worst case is $P(x_3|y=1) = 0.99999, P(x_3|y=0) = 0.0001$, then $\log\frac{P(x_3|y=1)}{P(x_3|y=0)}$ reach infinity.

however, the x are conditional independent. Thus the additional term's influence would be very small.

4. [15 points] **Conjugate priors for Gaussian distribution**

(a) [10 points] Consider a univariate Gaussian distribution $N(x|\mu, \lambda^{-1})$ having conjugate Gaussian-gamma prior given by

$$N(\mu|\mu_0, (\beta\lambda)^{-1})\text{Gam}(\lambda|a, b)$$

(see equation (2.154) in the Bishop textbook), where

$$\text{Gam}(\lambda|a, b) = \frac{1}{\Gamma(a)} b^a \lambda^{a-1} \exp(-b\lambda),$$

and $\lambda = 1/\sigma^2$ is the precision. Let a data set $x = \{x_1, \ldots, x_N\}$ be N i.i.d. observations. Show that the posterior distribution is also a Gaussian-gamma distribution of the same functional form as the prior, and write down expressions for the parameters of this posterior distribution.

(b) [5 points] Verify that the Wishart distribution

$$W(\mathbf{\Lambda}|\mathbf{W}, \nu) = B|\mathbf{\Lambda}|^{(\nu-D-1)/2} \exp\left(-\frac{1}{2} Tr(\mathbf{W}^{-1}\mathbf{\Lambda})\right)$$

(also defined by (2.155)) is indeed a conjugate prior for the precision matrix of a D-dimensional multivariate Gaussian, where $\mathbf{\Lambda}$ is an unknown precision matrix, $\mathbf{W}$ is a scale matrix, ν is the number of degrees of freedom, and Tr denotes the trace.

5. [40 points] Using discriminative vs. generative classifiers

For this problem, you will experiment with a version of the Wisconsin data set that we use as an illustration in class. The data is available in files wpbcx.dat and wpbcy.dat.

(a) [10 points] Implement logistic regression. If you use a learning-rate version, you will need to set up your code in such a way as to be able to search for a good learning rate. You can also use the iterative recursive least-squares version (whichever you prefer).

(b) [10 points] In a first experiment, use just a bias term and the first feature (first column in the wpbcx.dat _le). Set up a Gaussian naive Bayes classifier, and compare its results with logistic regression, using 10-fold cross-validation. Comment on what you observe.