COMP 4670/8600: Introduction to Statistical Machine Learning

Assignment I

Manab Chetia, u5492350

1. **Probabilities**
   1. **(2/20) Medical Test**
2. 0.078
3. Probability of disease, P(D) = 0.01

Probability that test is positive given person is diseased, P(+|D) = 0.8

Probability that test is positive given person is healthy, P(+|D’) = 0.095

Probability that person is diseased given test is positive, P(D|+) = ?

Using Bayes Theorem,

P(D|+) =

=

=

= 0.078

1. A person who is diseased but is detected healthy has a dangerous consequence, as there is a possibility the person spreads the disease to others; therefore it has the highest penalty of 1000.

A person who is healthy but is detected diseased is not much of a worry as there lies no threat of spreading the disease but there lies a low trouble as he has to undergo medical procedures which will take away precious time and effort. Hence a penalty of 1.

A person who is diseased and is detected diseased and a person who is healthy and is detected healthy are correctly diagnosed. Hence there are no penalties for both cases.

1. 2.09405
2. a) Diseased

b) Diseased

Let’s consider classes and .

Calculating losses,

As < , therefore we choose

* 1. **(2/20) Maximum Likelihood (ML) and Maximum A Posteriori (MAP)**

1. From Bayes’ theorem we know,

Therefore we can write,

1. We know,

As the i.i.d data have been already observed,

Therefore

4.

As is constant therefore,

* 1. **(3/20) Laplace Approximation**

1.

Mode

* Calculating second derivative of at we get,
* Calculating precision A,

As , therefore

As second derivative of at is negative as well as precision , Gaussian Approximation will be defined.

1. Mean = ,

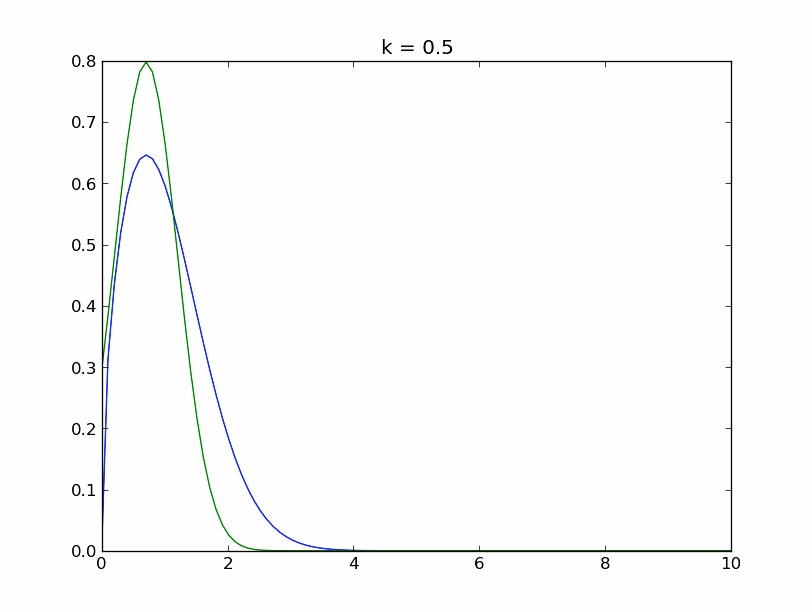
Variance =

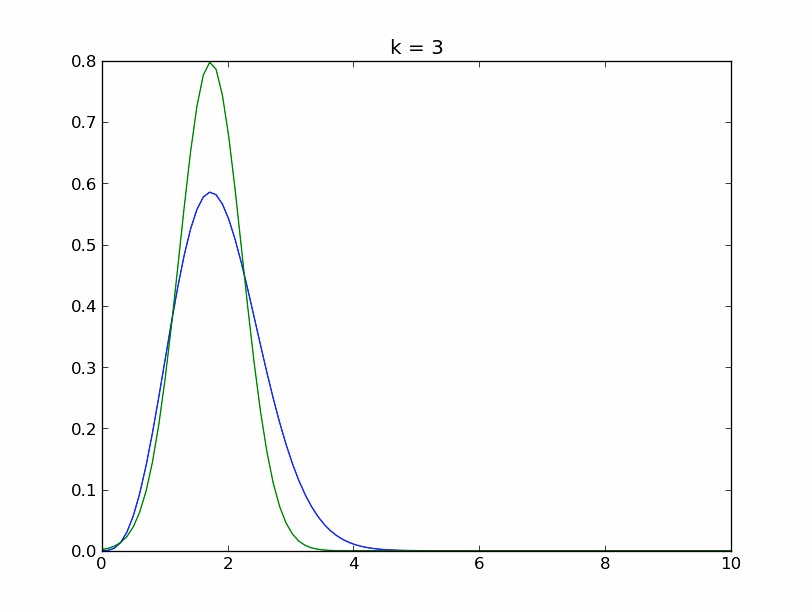
1. *k* = 0.5,

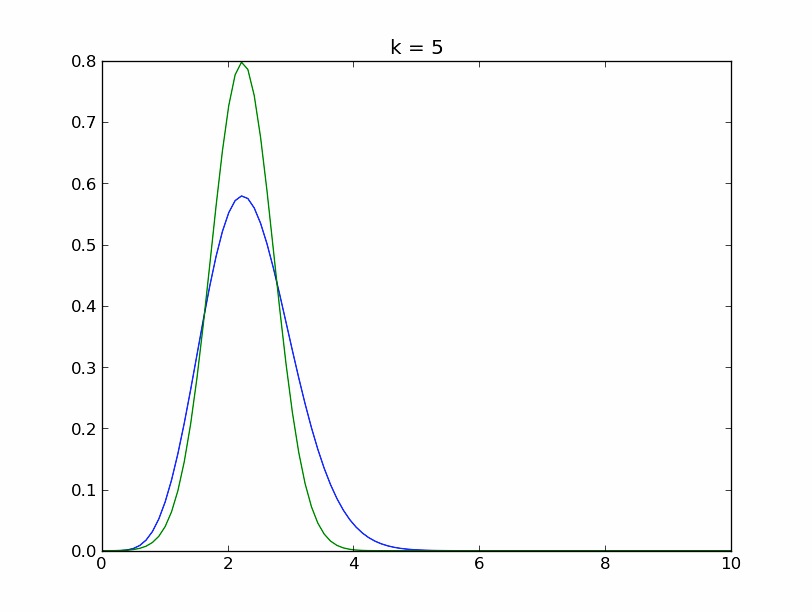
*k* = 3,

*k* = 5,

1. is a steeply decreasing function as increases. As increases, most values will be .
2. Plots







* 1. **(3/20) Bayesian Linear Regression**

1.

Multiplying both sides by we get

Considering only RHS,

=

=

=

=

=

=

=

Therefore L.H.S = R.H.S

1. We are given = where

Let’s assume variance of noise σ2v = and A =

Therefore,

… (a)

ϕ’ϕ

So we can write

… (b)

To compute we use,

We have a set of data set of N data points and adding a new data point (xN+1, tN+1).

Rewriting for points we get,

Therefore inverse of ,

… (c)

Substituting (c) in (b), we get,

… (d)

From (d), we can write

1. This inequality represents if extra data points are added, variance of output predictions will never be larger than any subset of those points.

**2. Dimensionality Reduction**

**2.1 Projection with Fisher's Discriminant**

1. [ [-0.00648404 0.0002029 0.01656621 0.00464915]

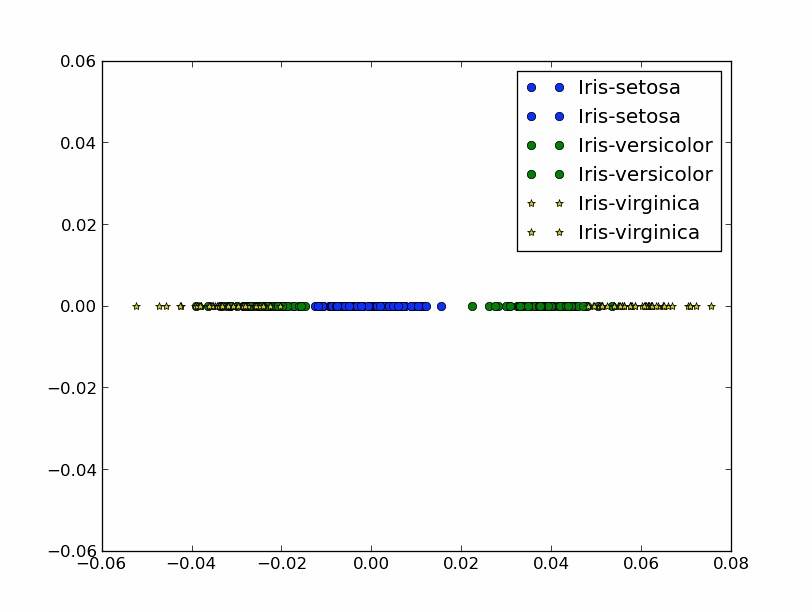
[-0.01199644 0.01822157 0.00289538 -0.0118346 ]

[ 0.01720897 -0.00784518 0.00506389 -0.01386845]

[ 0.02197205 0.02390111 -0.02562052 0.02471579] ]

2.a

,

2.b 

2.c The largest eigenvalue is 32.19, which is more than 110 times bigger than the next largest eigenvalue of 0.289. When projecting the 4 dimensional data into a 2 dimensional space defined by the eigenvectors (related with the 2 largest eigenvalues) we will get the maximum separation between the classes when projected in this 2D plane. However, the hyper plane associated with largest eigenvector () will separate the data more than the hyper plane associated with eigenvector .289.

4. for is

Same as 2.2.b

|  |  |  |  |
| --- | --- | --- | --- |
| **Planes** | **Associated Eigenvalues** | | **J** |
| **1** | 32.19 | 0.285 | 32.477 |
| **2** | 32.19 | -1.80e-15 | 32.191 |
| **3** | 32.19 | 7.62e-15 | 32.191 |
| **4** | 0.285 | -1.80e-15 | 0.285 |
| **5** | 0.285 | 7.62e-15 | 0.285 |
| **6** | -1.80e-15 | 7.62e-15 | 5.82e-15 |

From this table, we can see J associated with the largest eigenvalues are the maximum and the eigenvectors related with the largest J provides the maximum separation between the classes.

**3. Cross Validation and Classification**

3.

|  |  |
| --- | --- |
| **S** | ***k* at which minimum errors occurs** |
| 2 | 1 |
| 5 | 25 |
| 10 | 11 |

In case of several *k* having the same lowest error, we choose the largest of those *k* as we are averaging over more neighbors and the effective decision boundary is more smoother which prevents over-fitting.

4. As *S* increases, for each *k*, misclassification error decreases.

As *S* increases, number of samples in training set increases and number of samples for validation set decreases.

As a result, when *S* is large, more training data is there to give rise to a better generalization model. Also as size of cross validation set decreases, the model better estimates the cross validation points. Hence, fewer errors as S increase.

5. Optimal *k* increases at first and then decreases as *S* increases.

First, when *S* is small, we have less training data. When we have less training data, less number of points are scattered over the entire space. Therefore the optimal value of *k* is also less because:

* Number of neighbors is small.
* When a larger value of k is chosen, it leads to misclassification due to under-fitting of data.

Therefore minimum error occurs when we average over a small set of neighbors. Therefore we have a smaller optimal *k* when *S* is small.

As *S* increases, size of training data set increases, therefore giving rise to much more prominent neighbors. Consequently now we can average over a higher number of neighbors for a better generalization. As a result optimal *k* increases.

As *S* increases further, size of cross validation set reduces and size of training set increases. With the further increase in *S*, there are more number of neighbors to calculate the distance metric, leading to higher generalization. As the number of neighbors increases, the optimal value of *k* decreases because availability of more number of training points results in better estimation with lesser value of *k*.

6. Python code filename: *u5492350\_ManabChetia\_kNN.py*