Session 5
Classification Methods II

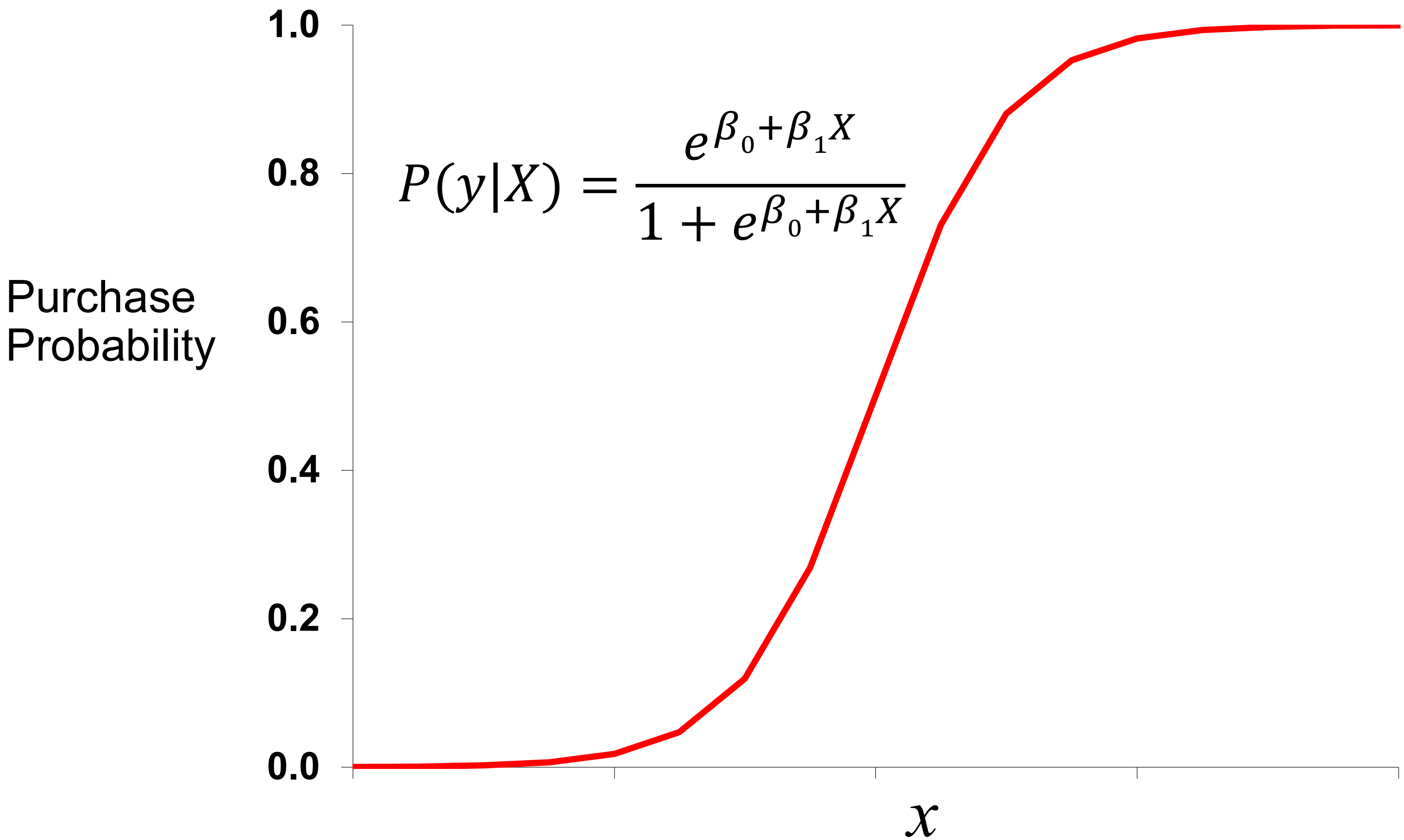
PSYC 560
Heungsun Hwang

Classification

- Classification is a process of predicting a nominal variable with multiple response categories, classes or labels
  - Assigns an observation to a category

- Popular classification methods or *classifiers*:
  - Logistic regression
  - Discriminant analysis
  - Naïve Bayes
  - K-nearest neighbors

Logistic Regression



$$P(y|X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$$

Estimating Probabilities

- Once the coefficients are estimated, it is simple to compute the probability of y = 1 for any given X values in a training or test sample. For example,

$$\hat{P}(y|X) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 X}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 X}}$$

*BBB_training.sav [DataSet1] - IBM SPSS Statistics Data Editor

File Edit View Data Transform Analyze Graphs Utilities Extensions Window Help

Visible: 14 of 14 Varia

	acctnum	gender	last	book$	child	youth	cook	do_it	reference	art	geog	buyer	probability	binary_pred	var	var	var	var	var
1	10003	1	15	25.00	0	0	2	0	0	0	0	0	.01515	0					
2	10006	1	7	15.00	0	1	0	0	0	0	0	1	.04725	0					
3	10013	1	13	15.00	0	0	0	1	0	0	0	0	.01886	0					
4	10015	1	25	15.00	0	0	1	0	0	0	0	0	.00747	0					
5	10016	1	1	23.00	2	0	0	0	0	0	0	0	.06565	0					
6	10017	1	7	39.00	0	0	2	0	0	0	1	0	.05761	0					
7	10019	1	11	26.00	0	0	1	1	0	0	0	0	.01740	0					
8	10022	1	15	15.00	0	0	0	0	1	0	0	0	.03198	0					
9	10025	1	13	25.00	0	1	1	0	0	0	0	0	.02107	0					
10	10026	1	15	15.00	1	0	0	0	0	0	0	0	.02060	0					
11	10030	1	13	15.00	1	0	0	0	0	0	0	0	.02489	0					
12	10035	1	13	29.00	0	0	0	0	0	1	1	0	.16501	0					
13	10036	1	9	15.00	0	0	0	1	0	0	0	0	.02754	0					
14	10037	1	7	15.00	0	0	0	0	0	0	1	0	.09985	0					
15	10040	1	9	15.00	1	0	0	0	0	0	0	0	.03624	0					
16	10042	1	11	25.00	1	0	1	0	0	0	0	0	.02348	0					
17	10044	1	13	101.00	1	1	3	2	0	1	1	0	.02445	0					
18	10045	1	33	15.00	0	0	1	0	0	0	0	0	.00345	0					
19	10046	1	21	78.00	2	1	2	0	1	0	1	0	.01250	0					
20	10048	0	29	126.00	3	0	2	2	0	3	1	0	.10771	0					
21	10049	1	11	26.00	1	0	0	1	0	0	0	0	.01845	0					
22	10051	1	9	15.00	0	0	1	0	0	0	0	0	.03422	0					
23	10052	1	21	23.00	2	0	0	0	0	0	0	0	.01003	0					
24	10054	1	17	27.00	0	0	1	0	1	0	0	0	.01981	0					
25	10055	1	5	137.00	2	1	3	0	1	1	4	1	.51409	1					
26	10058	0	3	59.00	1	1	1	1	0	0	1	0	.10203	0					
27	10062	1	13	15.00	0	0	0	0	0	1	0	0	.09738	0					
28	10063	0	13	67.00	1	0	3	1	0	1	0	0	.05013	0					
29	10064	1	1	29.00	0	1	0	0	0	0	1	0	.13973	0					

Data View Variable View

IBM SPSS Statistics Processor is ready Unicode:ON

Example: Logistic Regression

- The BookBinder Book Club data (BBB_training.csv & BBB_test.csv)
  - DV:
    - Buyer: Bought "Art History of Florence?"
  - Predictors:
    - Gender: 0 = male, 1 = female
    - Last : Months since last purchase
    - Book: Total $ spent on books
    - Art: # purchases of Art books
    - Child: # purchases of Children's books
    - Youth: # purchases of Youth books
    - Cook: # purchases of Cookbooks
    - Do_it: # purchases of Do-it-yourself books
    - Reference: # purchases of Reference books
    - Geog: # purchases of Geography books

Example: Logistic Regression

Training sample

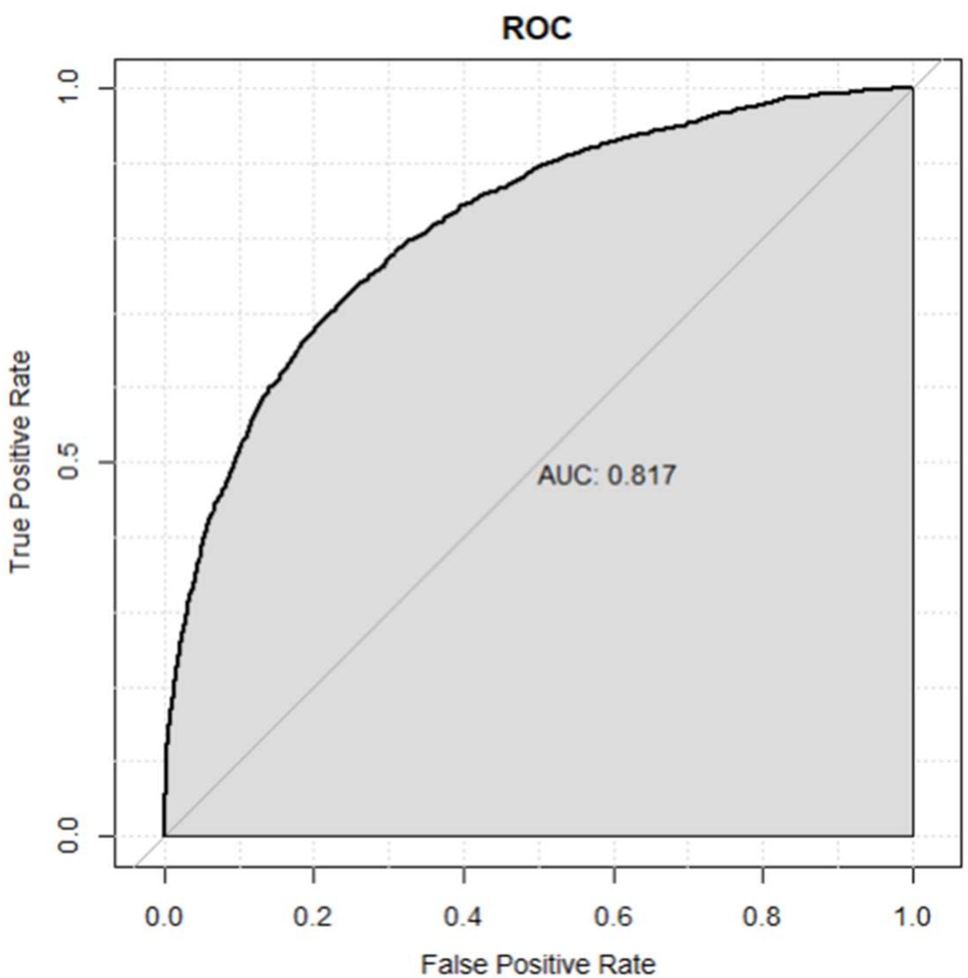
```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 22736  1901
         1   190   368

               Accuracy : 0.917
                 95% CI : (0.9135, 0.9204)
    No Information Rate : 0.9099
    P-Value [Acc > NIR] : 3.894e-05

                  Kappa : 0.2331

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.16219
            Specificity : 0.99171
         Pos Pred Value : 0.65950
         Neg Pred Value : 0.92284
             Prevalence : 0.09006
         Detection Rate : 0.01461
   Detection Prevalence : 0.02215
      Balanced Accuracy : 0.57695

       'Positive' Class : 1
```

Test sample

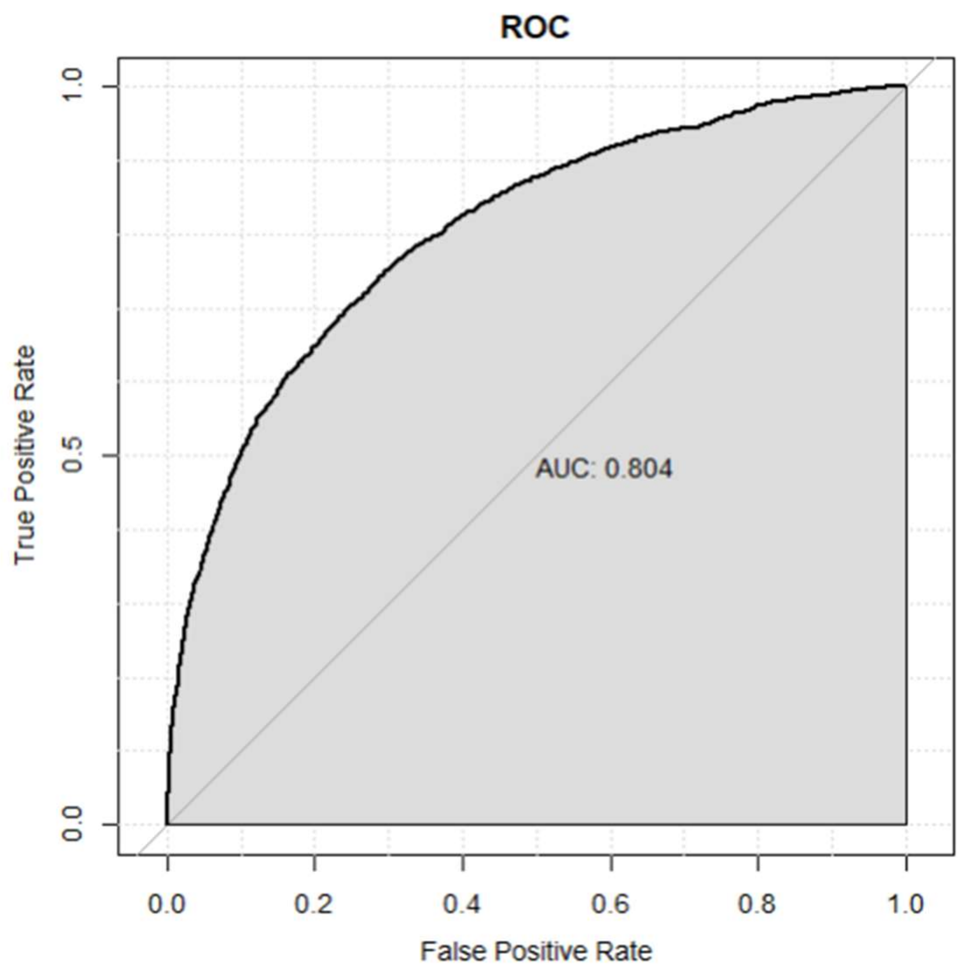
```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 22365  1921
         1   187   332

               Accuracy : 0.915
                 95% CI : (0.9115, 0.9185)
    No Information Rate : 0.9092
    P-Value [Acc > NIR] : 0.0006381

                  Kappa : 0.2128

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.14736
            Specificity : 0.99171
         Pos Pred Value : 0.63969
         Neg Pred Value : 0.92090
             Prevalence : 0.09083
         Detection Rate : 0.01338
   Detection Prevalence : 0.02092
      Balanced Accuracy : 0.56953

       'Positive' Class : 1
```

Example: Logistic Regression

Training sample



Test sample



Other Classification Methods

- Logistic regression involves directly modeling P(Y|X) using the logistic function.
  - We model the conditional probability of Y given X.

- We now consider alternative and less direct methods for estimating these probabilities.
  - We model the distributions of predictors (X) separately in each of the response classes. We then use Bayes' theorem to flip these around into estimates for P(Y|X).

Other Classification Methods

- Why consider another classifier over logistic regression?

  - If the sample size is small and the distribution of X is approximately normal per class, the alternative methods are more stable than logistic regression.

  - The alternative methods are popular when we have more than 2 response classes.

Other Classification Methods

- Suppose that we wish to classify an observation into one of K classes ($K \geq 2$).

- π_k = overall or prior probability that an observation comes from the kth class
  - ≈ class size

- $f_k(X) = P(X|Y = k)$: The density function of X for an observation that comes from the kth class

Other Classification Methods

- Then, Bayes' theorem states that

$$P(Y = k|X = x) = \frac{\pi_k f_k(X)}{\sum_{l=1}^{K} \pi_l f_l(X)}. \qquad \text{Eq.(1)}$$

- $P(Y = k|X = x)$ is called the posterior probability that an observation belongs to the *k*th class given the predictor value for the observation.

Other Classification Methods

- Instead of directly computing the posterior probability as in logistic regression, we can plug in estimates of π_k and $f_k(X)$ into Eq. (1).

- In general, it is easy to estimate π_k (the proportion of the training observations that belong to the *k*th class). Yet, estimating $f_k(X)$ is more difficult.

Other Classification Methods

- We discuss three classifiers that use different estimates of $f_k(X)$.
  - Linear discriminant analysis (LDA)
  - Quadratic discriminant analysis (QDA)
  - Naïve Bayes

Linear Discriminant Analysis with One Predictor

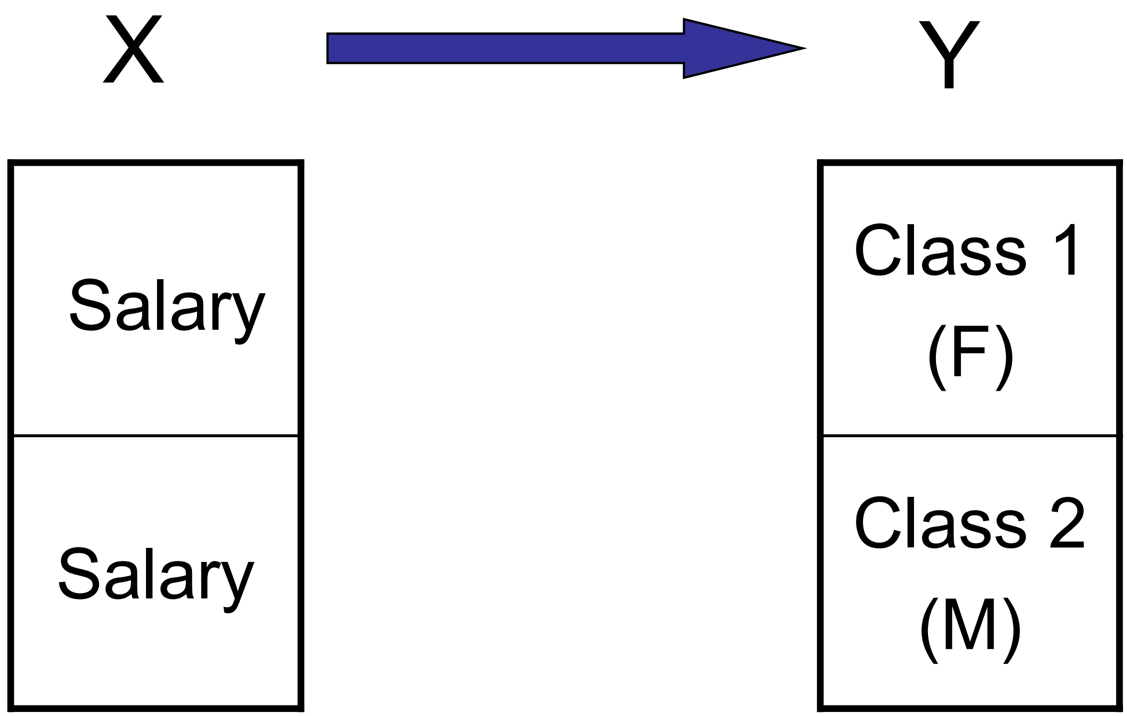
- Assume that we have only one predictor. Our task is to estimate $f_k(X)$ (and estimate Eq. (1)).

- In LDA, we assume that $f_k(X)$ is normal or Gaussian. When P = 1, the normal density takes the form
$$f_k(X) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{1}{2\sigma_k^2}(X-\mu_k)^2\right). \quad \text{Eq. (2)}$$

- We further assume that all variances are the same across K classes ($\sigma_1^2 = \ldots = \sigma_K^2$).

- $X|Y = k \sim N(\mu_k, \sigma^2)$

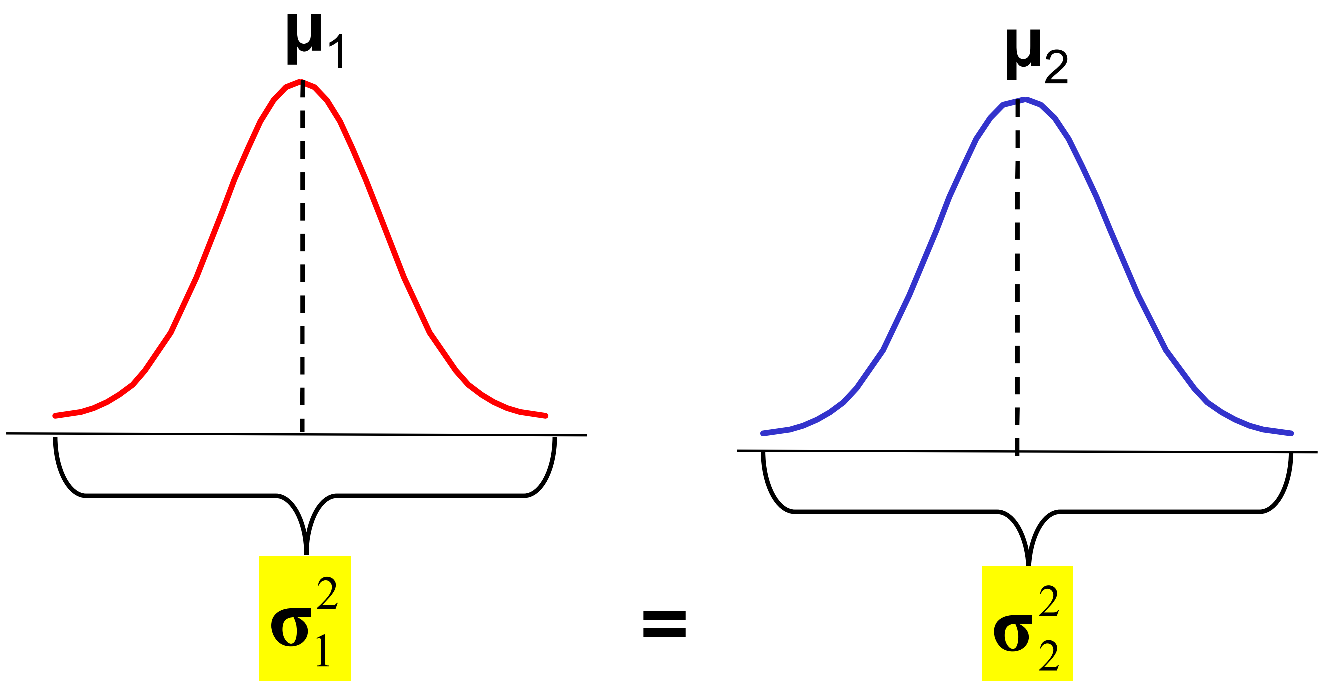
Linear Discriminant Analysis with One Predictor



Linear Discriminant Analysis with One Predictor

$X|Y = 1, 2$

Salary | Y = F	Salary | Y = M



$X|Y = 1 \sim N(\mu_1, \sigma^2)$ $\quad$ $X|Y = 2 \sim N(\mu_2, \sigma^2)$

Linear Discriminant Analysis with One Predictor

- Plugging Eq. (2) into Eq. (1), we find that

$$P(Y = k|X = x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{1}{2\sigma_k^2}(x-\mu_k)^2\right)}{\sum_{l=1}^{K} \pi_l \frac{1}{\sqrt{2\pi}\sigma_l} \exp\left(-\frac{1}{2\sigma_l^2}(x-\mu_l)^2\right)}. \quad \text{Eq. (3)}$$

- We can classify an observation $X = x$ to the class for which Eq. (3) is largest.

Linear Discriminant Analysis with One Predictor

- Taking the log of Eq. (3) and rearranging the terms, this is equivalent to assigning the observation to the class for which

$$\delta_k(x) = x\frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k)$$

 is largest.

- In LDA, we estimate μ_k as the average of all the training observations from the *k*th class; σ^2 as a weighed average of the sample variances for *K* classes; and π_k as the proportion of the training observations that belong to the *k*th class.

Linear Discriminant Analysis with One Predictor

- The LDA classifier assigns an observation $X = x$ to the class for which
$$\hat{\delta}_k(x) = x\frac{\hat{\mu}_k}{\hat{\sigma}^2} - \frac{\hat{\mu}_k^2}{2\hat{\sigma}^2} + \log(\hat{\pi}_k)$$
is largest.

- The word "linear" in LDA stems from the fact that the discriminant functions $\hat{\delta}_k(x)$ are linear functions of x.

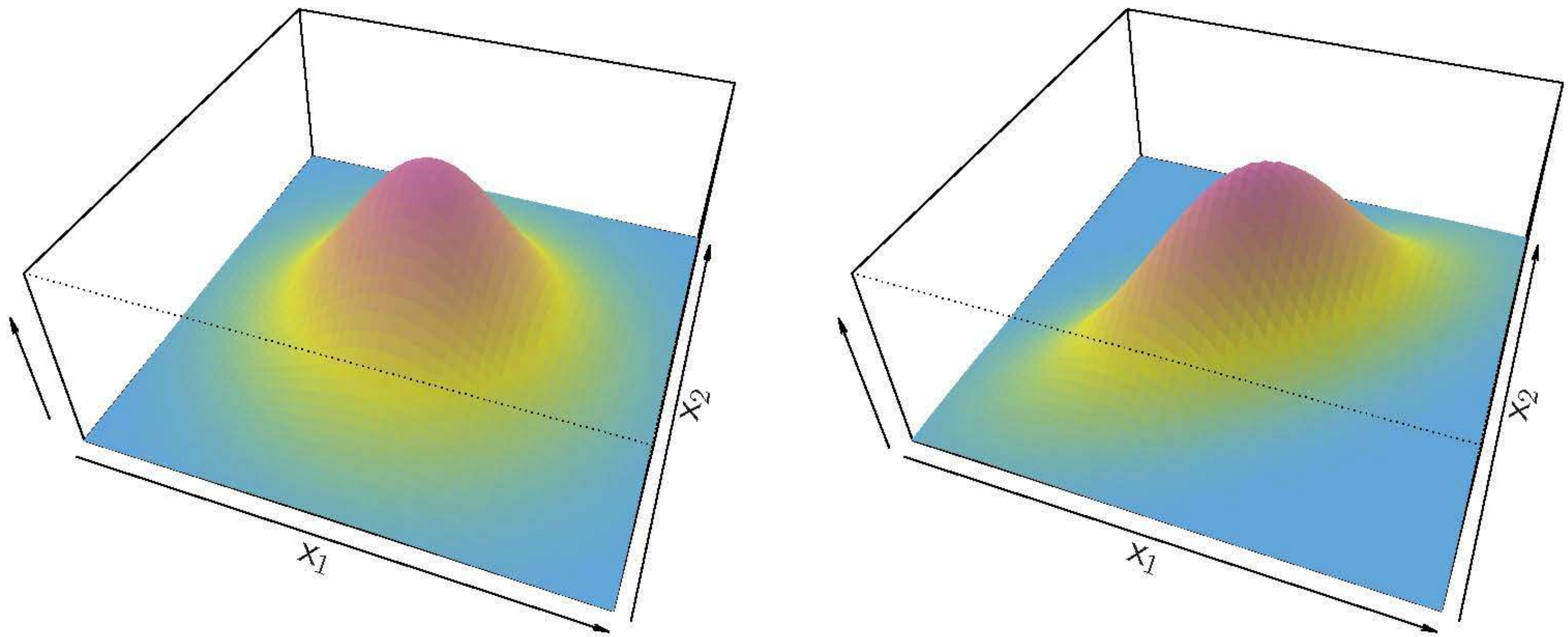
Linear Discriminant Analysis with Multiple Predictors

- In the case of P > 1 predictors, the LDA classifier assumes that the observations in the *k*th class are drawn from a multivariate normal distribution with P means and the covariance matrix of the predictors.

- The covariance matrix is assumed to be common to all *K* classes.
  - $\boldsymbol{X} | Y = k \sim N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$

Linear Discriminant Analysis with Multiple Predictors

Multivariate normal distributions



James et al., (2021). Figure 4.5: Left: X1 and X2 are uncorrelated. Right: X1 and X2 are correlated (r = .7)

Linear Discriminant Analysis with Multiple Predictors

- The LDA classifier assigns an observation $X = \boldsymbol{x}$ to the class for which

$$\hat{\delta}_k(\boldsymbol{x}) = \boldsymbol{x}^T \widehat{\boldsymbol{\Sigma}}^{-1} \widehat{\boldsymbol{\mu}}_k - \frac{1}{2} \widehat{\boldsymbol{\mu}}_k^T \widehat{\boldsymbol{\Sigma}}^{-1} \widehat{\boldsymbol{\mu}}_k + \log(\hat{\pi}_k)$$

  is largest.

- Again, the discriminant function $\hat{\delta}_k(\boldsymbol{x})$ is a linear function of $\boldsymbol{x}$.

Quadratic Discriminant Analysis

- Like LDA, quadratic discriminant analysis (QDA) assumes that the observations in the *k*th class are drawn from a multivariate normal distribution.

- Unlike LDA, QDA assumes that each class has its own covariance matrix.
  - $\boldsymbol{x} | \, Y = k \sim N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$

Quadratic Discriminant Analysis

- This leads the discriminant function to be a quadratic function of $\boldsymbol{x}$.

$$\hat{\delta}_k(\boldsymbol{x}) = -\frac{1}{2}\boldsymbol{x}^T\widehat{\boldsymbol{\Sigma}}_k^{-1}\boldsymbol{x} + \boldsymbol{x}^T\widehat{\boldsymbol{\Sigma}}_k^{-1}\widehat{\boldsymbol{\mu}}_k - \frac{1}{2}\widehat{\boldsymbol{\mu}}_k^T\widehat{\boldsymbol{\Sigma}}_k^{-1}\widehat{\boldsymbol{\mu}}_k - \frac{1}{2}\log|\widehat{\boldsymbol{\Sigma}}_k| + \log(\hat{\pi}_k)$$

From $\hat{\delta}_k(\boldsymbol{x})$ to Probabilities

- Once we have estimates $\hat{\delta}_k(\boldsymbol{X})$, we can turn these into estimates for posterior probabilities:

$$\hat{P}(Y = k|X = x) = \frac{e^{\hat{\delta}_k(\boldsymbol{x})}}{\sum_{l=1}^{K} e^{\hat{\delta}_l(\boldsymbol{x})}}.$$

- Thus, classifying to the largest $\hat{\delta}_k(\boldsymbol{X})$ is equivalent to classifying to the class for which $\hat{P}(Y = k|X = x)$ is largest.

LDA VS. QDA

- Why would one prefer LDA to QDA or vice-versa?
- The answer lies in the bias-variance trade-off.
  - QDA involves much more parameters (i.e., each class's covariances) than LDA.
  - LDA is a much less flexible classifier than QDA, so has a lower variance. This may improve prediction.
  - However, if LDA's assumption of a common covariance matrix for K classes is badly off, it can suffer from higher bias.

LDA VS. QDA

- Roughly speaking, LDA may be chosen over QDA if the number of training observations is relatively small.

- QDA may be recommended if the number of training observations is very large or if the assumption of a common covariance matrix for K classes is clearly untenable.

Example: Linear Discriminant Analysis

Training sample

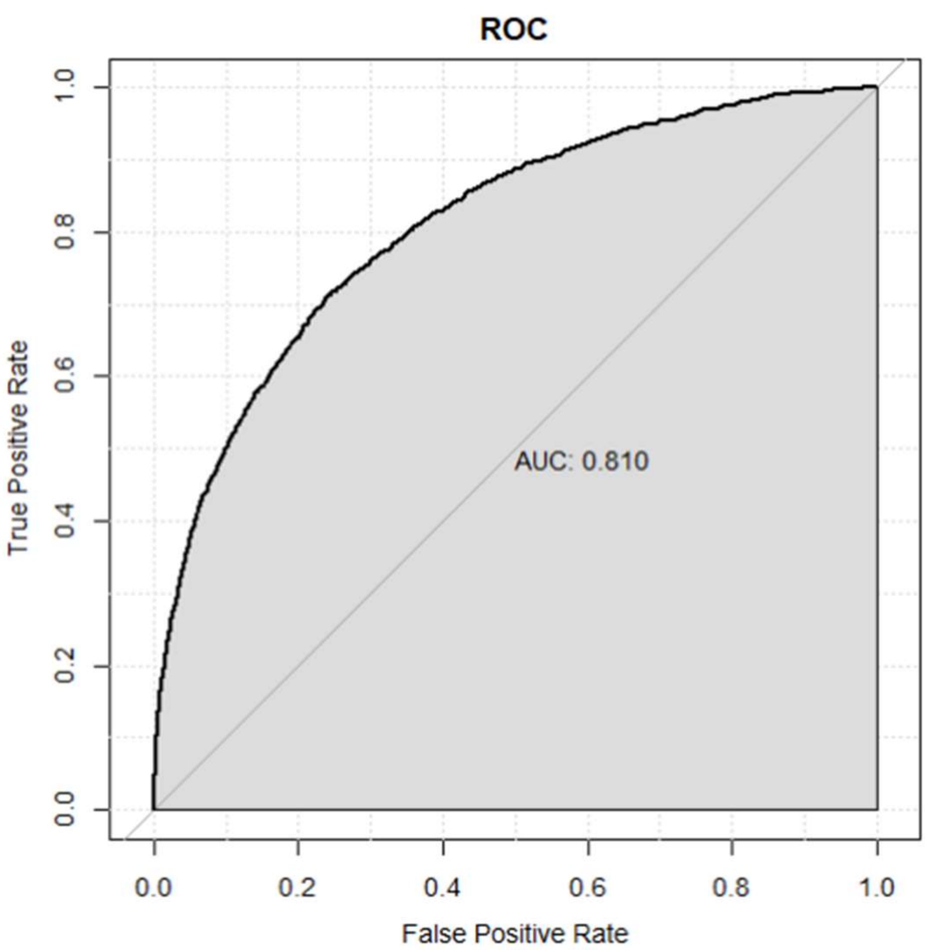
```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 22455  1718
         1   471   551

               Accuracy : 0.9131
                 95% CI : (0.9096, 0.9166)
    No Information Rate : 0.9099
    P-Value [Acc > NIR] : 0.03955

                  Kappa : 0.2954

 Mcnemar's Test P-Value : < 2e-16

            Sensitivity : 0.24284
            Specificity : 0.97946
         Pos Pred Value : 0.53914
         Neg Pred Value : 0.92893
             Prevalence : 0.09006
         Detection Rate : 0.02187
   Detection Prevalence : 0.04056
      Balanced Accuracy : 0.61115

       'Positive' Class : 1
```

Test sample

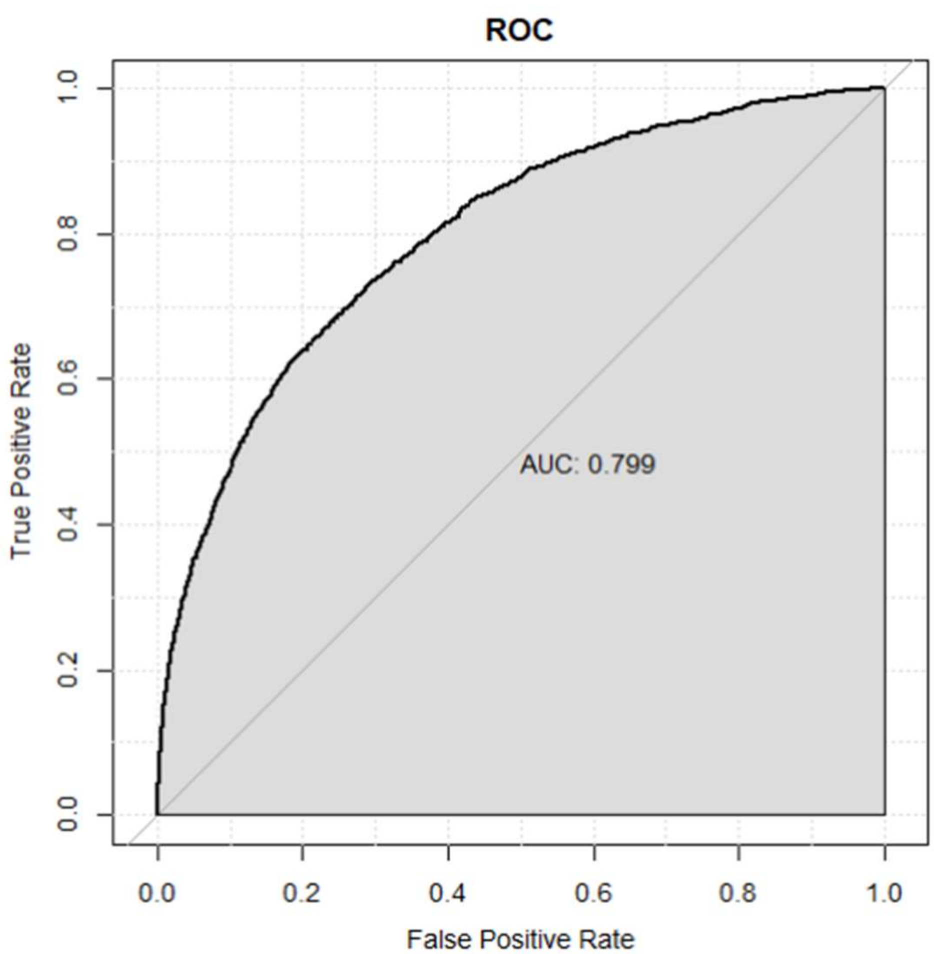
```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 22084  1731
         1   468   522

               Accuracy : 0.9113
                 95% CI : (0.9077, 0.9149)
    No Information Rate : 0.9092
    P-Value [Acc > NIR] : 0.1183

                  Kappa : 0.2821

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.23169
            Specificity : 0.97925
         Pos Pred Value : 0.52727
         Neg Pred Value : 0.92731
             Prevalence : 0.09083
         Detection Rate : 0.02104
   Detection Prevalence : 0.03991
      Balanced Accuracy : 0.60547

       'Positive' Class : 1
```

Example: Linear Discriminant Analysis

Training sample



Test sample



Example: Quadratic Discriminant Analysis

Training sample

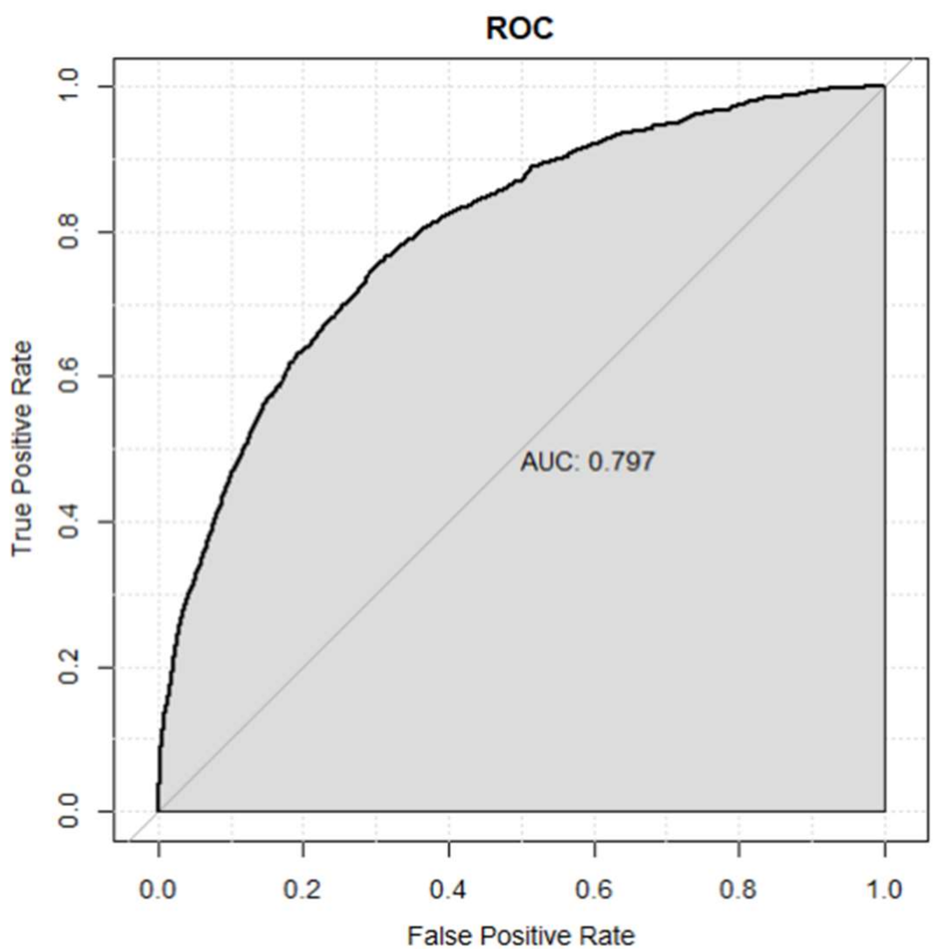
```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 21738  1533
         1  1188   736

               Accuracy : 0.892
                 95% CI : (0.8881, 0.8958)
    No Information Rate : 0.9099
    P-Value [Acc > NIR] : 1

                  Kappa : 0.2926

 Mcnemar's Test P-Value : 4.262e-11

            Sensitivity : 0.32437
            Specificity : 0.94818
         Pos Pred Value : 0.38254
         Neg Pred Value : 0.93412
             Prevalence : 0.09006
         Detection Rate : 0.02921
   Detection Prevalence : 0.07636
      Balanced Accuracy : 0.63628

       'Positive' Class : 1
```

Test sample

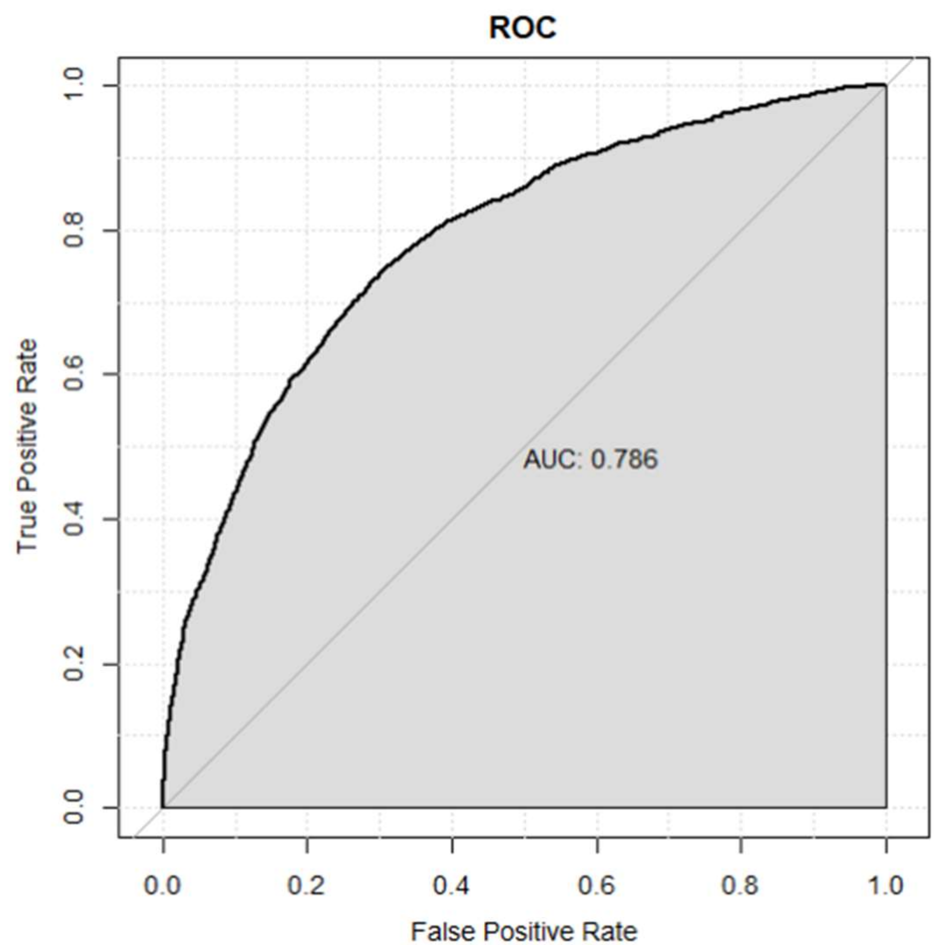
```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 21377  1559
         1  1175   694

               Accuracy : 0.8898
                 95% CI : (0.8858, 0.8937)
    No Information Rate : 0.9092
    P-Value [Acc > NIR] : 1

                  Kappa : 0.2772

 Mcnemar's Test P-Value : 2.391e-13

            Sensitivity : 0.30803
            Specificity : 0.94790
         Pos Pred Value : 0.37132
         Neg Pred Value : 0.93203
             Prevalence : 0.09083
         Detection Rate : 0.02798
   Detection Prevalence : 0.07535
      Balanced Accuracy : 0.62797

       'Positive' Class : 1
```

Example: Quadratic Discriminant Analysis

Training sample



Test sample



Naïve Bayes

- The naïve Bayes (also known as "Idiot's Bayes") classifier estimates $f_k(X)$ in a different way. Instead of assuming a particular family of distributions (e.g., multivariate normal) for the density function, it makes a single assumption: "Within the *k*th class, the P predictors are independent."

- Mathematically, this assumption means that for *k* = 1,…, *K*,

$$f_k(X) = f_{k1}(X_1) \times f_{k2}(X_2) \times \cdots \times f_{kp}(X_p) = \prod_{p=1}^{P} f_{kp}(X_p),$$

where f_{kp} is the density function of the *p*th predictor among observations in the *k*th class.

Naïve Bayes

- Why is this assumption so powerful? Essentially, estimating a P-dimensional density function is challenging because we must consider not only the marginal distribution of each predictor — that is, the distribution of each predictor on its own — but also the joint distribution of the predictors — that is, the association between the different predictors.
  - In the case of a multivariate normal distribution, the association between the different predictors is summarized by the off-diagonal elements of the covariance matrix.
- However, in general, this association can be very challenging to estimate.

Naïve Bayes

- By assuming that the P predictors are independent within each class, we eliminate the need to worry about the association between the predictors, because we have assumed that there is no association between the predictors. Thus, it can simplify the estimation of the density function dramatically.

- In most settings, the naïve Bayes assumption is rather too optimistic and is generally not true.

Naïve Bayes

- Nonetheless, although this assumption is made for convenience, it often leads to quite decent results, especially when the sample size is not large enough relative to the number of predictors to effectively estimate the joint distribution of the predictors within each class.

- In fact, since estimating a joint distribution requires such a huge amount of data, naïve Bayes is a good choice in a wide range of settings.

- The naïve Bayes assumption introduces some bias, but reduces variance, leading to a classifier that works quite well in practice as a result of the bias-variance trade-off.

Example: Naïve Bayes

Training sample

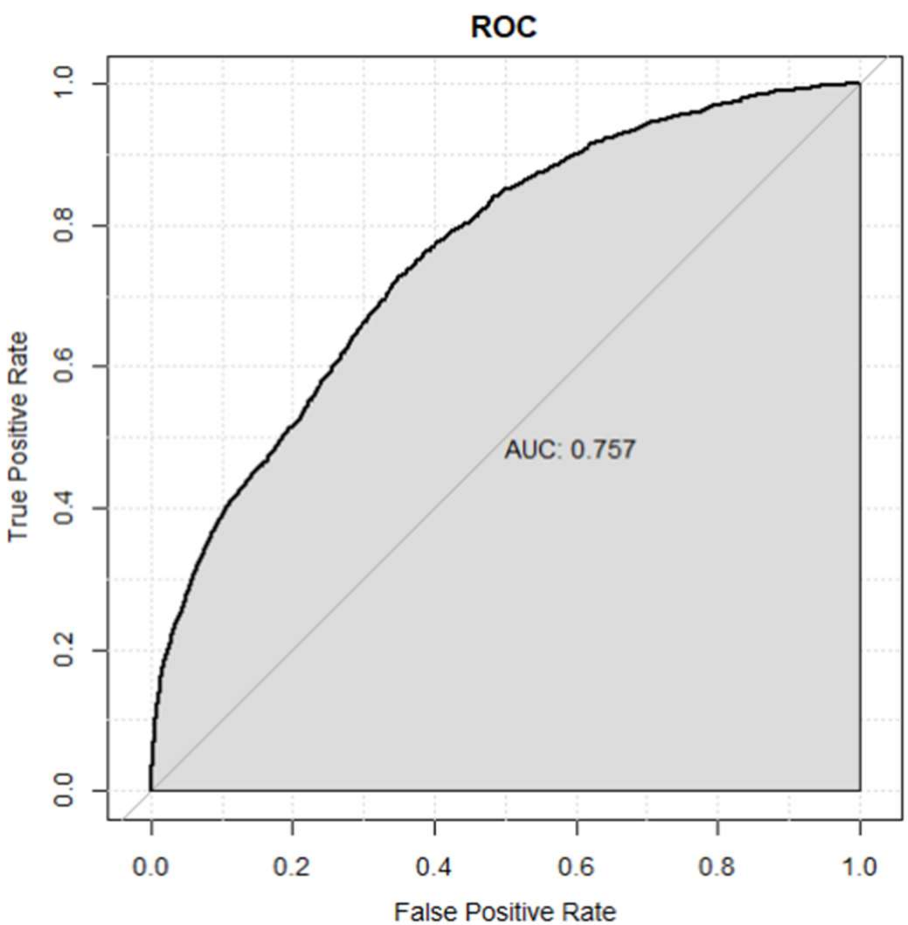
```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 20968  1446
         1  1958   823

               Accuracy : 0.8649
                 95% CI : (0.8606, 0.8691)
    No Information Rate : 0.9099
    P-Value [Acc > NIR] : 1

                  Kappa : 0.2517

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.36271
            Specificity : 0.91459
         Pos Pred Value : 0.29594
         Neg Pred Value : 0.93549
             Prevalence : 0.09006
         Detection Rate : 0.03267
   Detection Prevalence : 0.11038
      Balanced Accuracy : 0.63865

       'Positive' Class : 1
```

Test sample

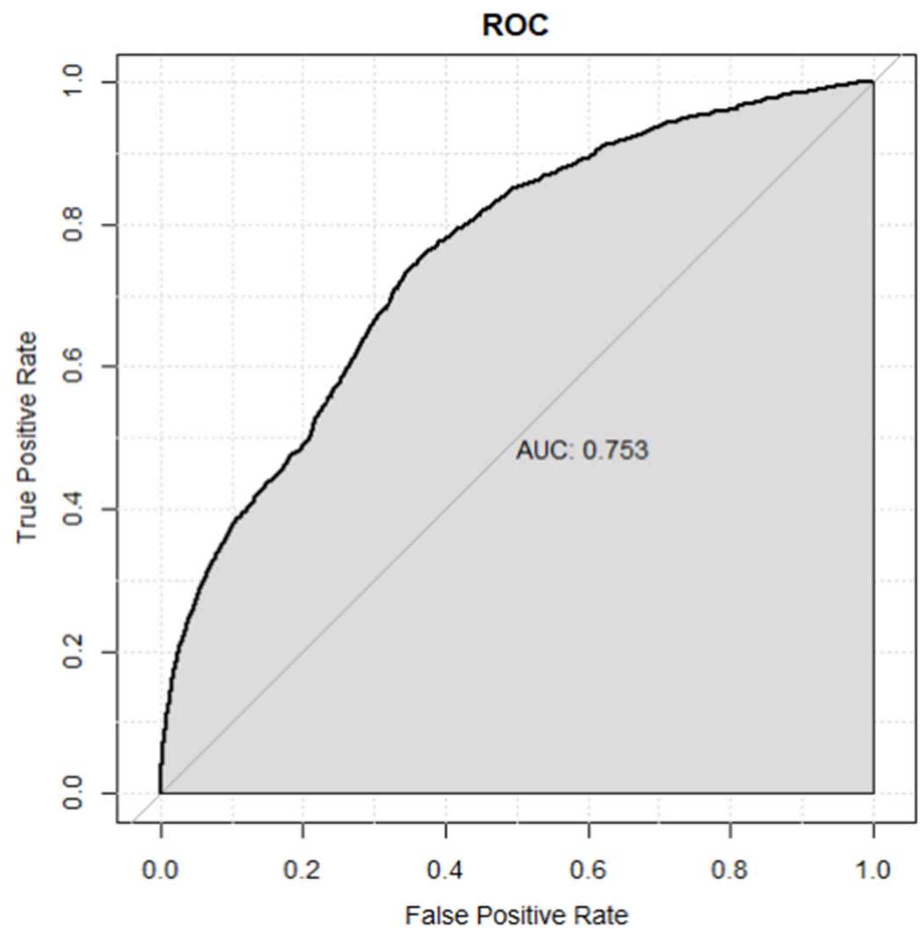
```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 20706  1481
         1  1846   772

               Accuracy : 0.8659
                 95% CI : (0.8616, 0.8701)
    No Information Rate : 0.9092
    P-Value [Acc > NIR] : 1

                  Kappa : 0.2431

 Mcnemar's Test P-Value : 2.778e-10

            Sensitivity : 0.34265
            Specificity : 0.91814
         Pos Pred Value : 0.29488
         Neg Pred Value : 0.93325
             Prevalence : 0.09083
         Detection Rate : 0.03112
   Detection Prevalence : 0.10554
      Balanced Accuracy : 0.63040

       'Positive' Class : 1
```

Example: Naïve Bayes

Training sample



Test sample



K-Nearest Neighbors (KNN)

- KNN is a completely non-parametric approach.
  - No assumptions are made about the shape of the decision boundary.
- Given a value for K and a prediction point x_0, KNN identifies the K points in the training data that are closest to x_0, represented by Q. It then estimates the conditional probability for class k as the fraction of points in Q whose response values equal to k:

$$P(Y = k|X = x_0) = \frac{1}{K} \sum_{x_i \in Q} I(y_i = k)$$

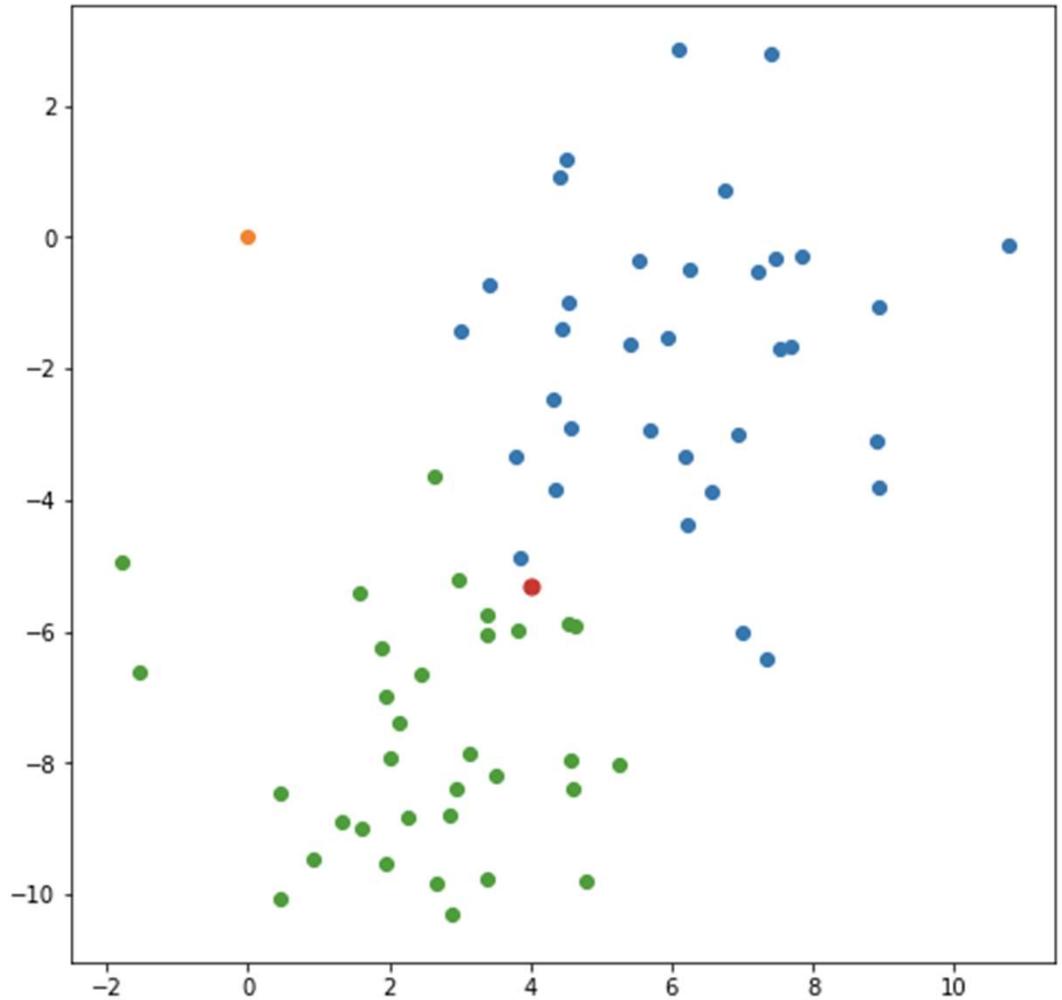
- KNN then classifies x_0 the class with the largest probability.

K-Nearest Neighbors (KNN)

- As in KNN regression, the choice of K has a drastic effect on the KNN classifier obtained.

- When K = 1, the decision boundary is overly flexible, resulting in low bias yet very high variance. As K increases, the method becomes less flexible and produces a decision boundary that is close to linear (so, high bias yet low variance).

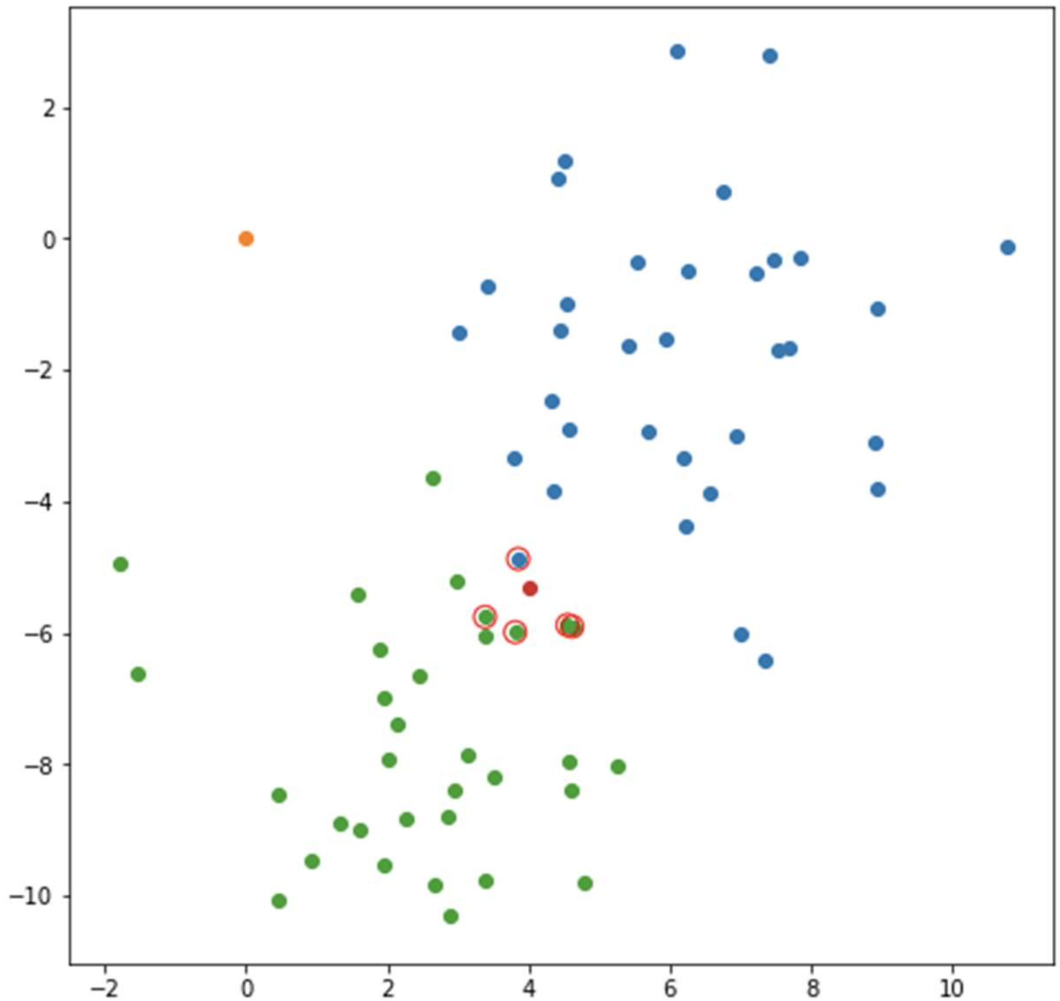
K-Nearest Neighbors (KNN)



https://towardsdatascience.com/the-basics-knn-for-classification-and-regression-c1e8a6c955

K-Nearest Neighbors (KNN)

K = 5



https://towardsdatascience.com/the-basics-knn-for-classification-and-regression-c1e8a6c955

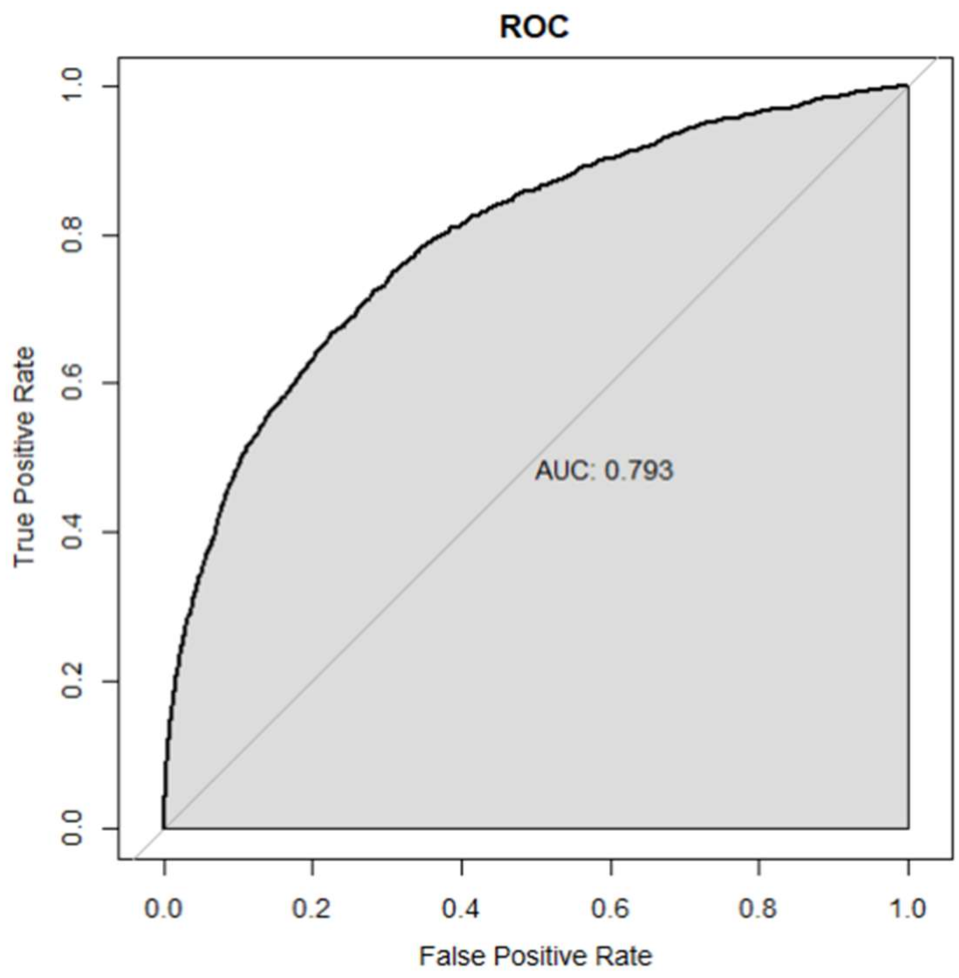
Example: K-Nearest Neighbors

Test sample

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 22541  2202
         1    11    51

               Accuracy : 0.9108
                 95% CI : (0.9072, 0.9143)
    No Information Rate : 0.9092
    P-Value [Acc > NIR] : 0.1916

                  Kappa : 0.0394

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.022636
            Specificity : 0.999512
         Pos Pred Value : 0.822581
         Neg Pred Value : 0.911005
             Prevalence : 0.090828
         Detection Rate : 0.002056
   Detection Prevalence : 0.002499
      Balanced Accuracy : 0.511074

       'Positive' Class : 1
```

Example: K-Nearest Neighbors

Test sample



A Comparison of Classification Methods

- Logistic regression can outperform LDA if LDA's assumptions (a normal distribution with the same covariance matrix) are violated.

- LDA can provide some improvements (lower variance) over logistic regression if the assumptions are met.

- However, in practice, the LDA assumptions are never correct. So, logistic regression seems to be a safer, more robust bet over LDA, relying on fewer assumptions (Hastie et al., 2009, p. 128).

A Comparison of Classification Methods

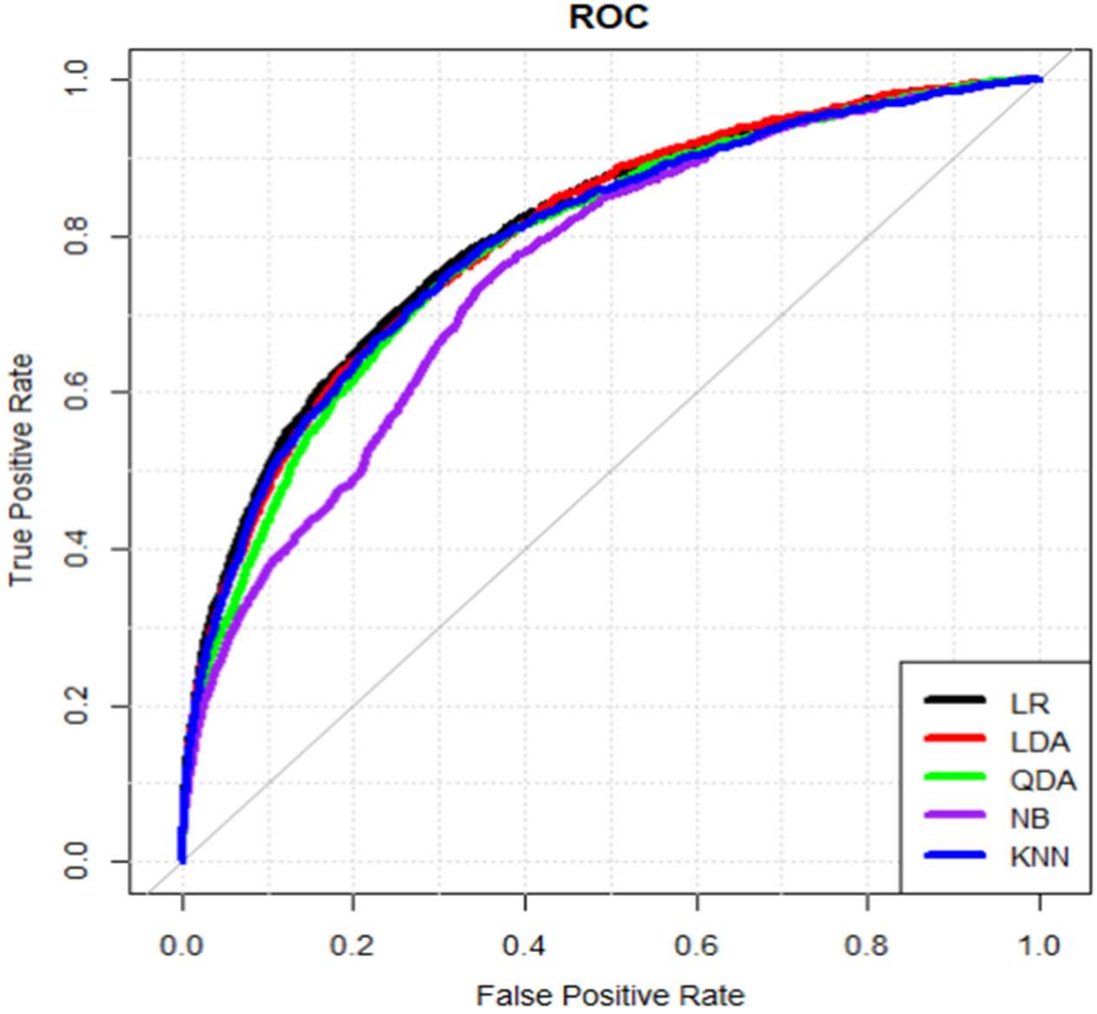
- For accurate classification, KNN requires a lot of observations relative to the number of predictors because it is non-parametric and tends to reduce the bias while incurring a lot of variance.

- KNN is expected to outperform LDA and logistic regression when the decision boundary is highly non-linear, provided that the sample size is very large and the number of predictors is small.

- When the decision boundary is non-linear but the sample size is only modest or the number of predictors is not very small, QDA may be preferred to KNN.

Example: A Comparison of Classification Methods

Test sample



AUC_LR:	0.804052
AUC_LDA:	0.7987326
AUC_QDA:	0.7862619
AUC_NB:	0.7526996
AUC_KNN:	0.7934951