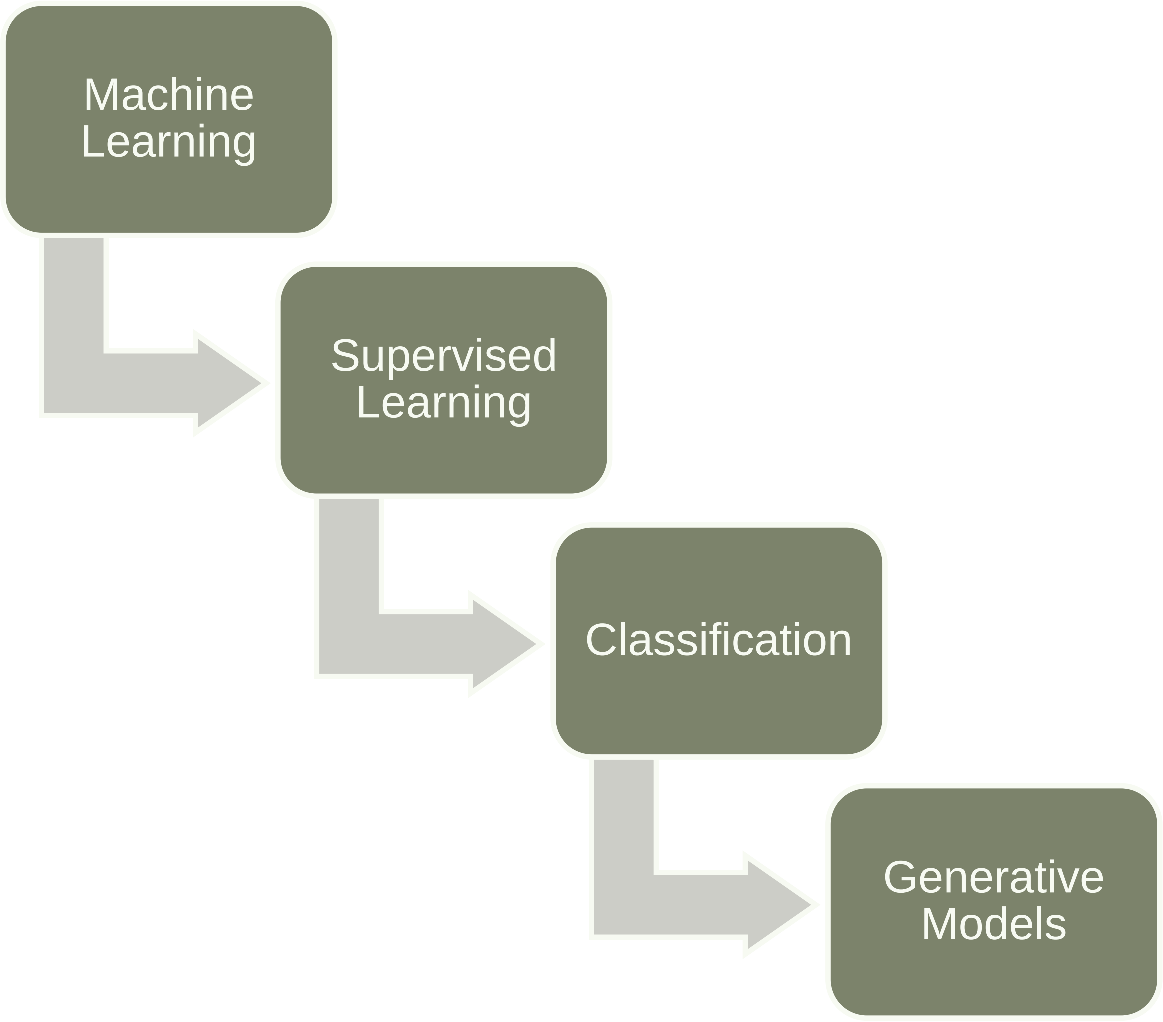
X4. Generative Models – Discriminant Analysis



Why do we need another method?

- When the classes of Y are substantially separated, the parameter estimates for the logistic regression model are unstable
- If n is small and the distribution of the predictors X is approximately normal in each of the classes, the linear discriminant model is again more stable than the logistic regression model
- It can be naturally extended to multiple classes (Y classes > 2)

The concept of discriminant analysis

- It is based on Bayes theorem: the conditional probability of Y happens given X can be calculated using the conditional probability of X happens given Y:

$$\Pr(Y = k|X = x) = \frac{\Pr(X = x|Y = k) \cdot \Pr(Y = k)}{\Pr(X = x)}$$

- Let's look at a simple example: If a lamp is defective, what is the probability that it's produced by Factory C, i.e., P(C|D)?

Factory	% of total production	Probability of defective lamps
A	$0.35 = P(A)$	$0.015 = P(D \mid A)$
B	$0.35 = P(B)$	$0.010 = P(D \mid B)$
C	$0.30 = P(C)$	$0.020 = P(D \mid C)$

$$P(C|D) = \frac{P(D|C) * P(C)}{P(D)} = \frac{0.02 * 0.3}{0.015 * 0.35 + 0.01 * 0.35 + 0.3 * 0.02} = 0.40678$$

We can write the formula as

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

- $f_k(x) = \Pr(X = x|Y = k)$ is the *density* for X in class k. Here we will use normal densities for these, separately in each class.
- $\pi_k = \Pr(Y = k)$ is the marginal or *prior* probability for class k.

Bayesian classifier

- We can calculate $p_k(X) = \Pr(Y = k|X)$ by plugging in π_k and $f_k(X)$
- In general, estimating *prior* probability π_k is easy: if we have random sample of *Y*s, just to compute the fraction of the observations that belong to *k*th class
- For $f_k(X)$, we need to assume some simple forms, e.g., *normal distribution*
- $p_k(X)$ is the *posterior* probability given the predictor value X, which is the probability we want to estimate
- We estimate a probability for each Y class given X and classify an observation to the class for which $p_k(X)$ is largest

How to estimate

- Usually, we do not know the parameters of our assumed distribution of X or prior probabilities
- We can use the data to estimate
- For example, with one variable X

$$
\begin{aligned}
\hat{\pi}_k &= \frac{n_k}{n} \\
\hat{\mu}_k &= \frac{1}{n_k} \sum_{i:\, y_i = k} x_i \\
\hat{\sigma}^2 &= \frac{1}{n-K} \sum_{k=1}^{K} \sum_{i:\, y_i = k} (x_i - \hat{\mu}_k)^2 \\
&= \sum_{k=1}^{K} \frac{n_k - 1}{n - K} \cdot \hat{\sigma}_k^2
\end{aligned}
$$

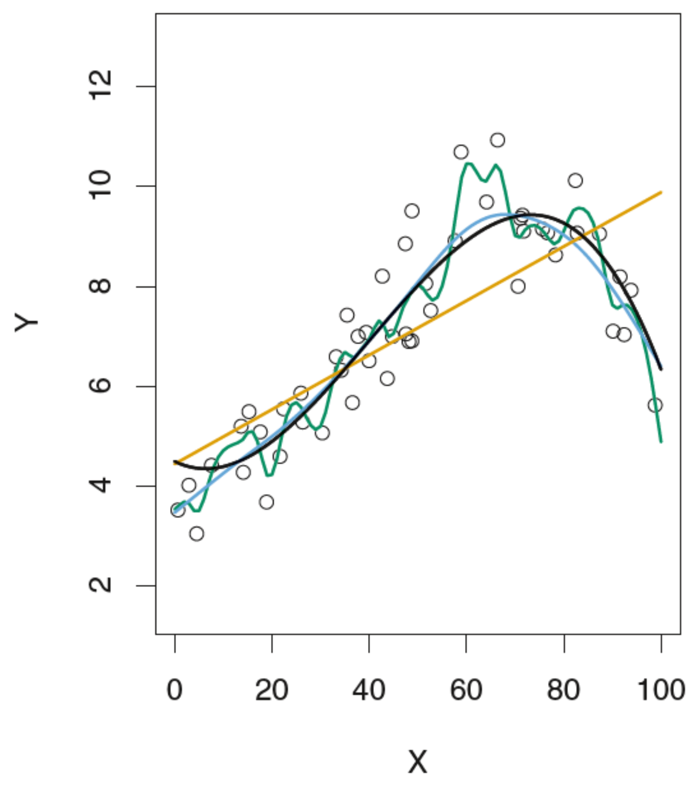
where $\hat{\sigma}_k^2 = \frac{1}{n_k - 1} \sum_{i:\, y_i = k} (x_i - \hat{\mu}_k)^2$ is the usual formula for the estimated variance in the kth class.

Different forms of Discriminant Analysis

- Based on different assumption about X distributions, we have different types of discriminant analysis
- Linear Discriminant Analysis (LDA)
  - Normal (Gaussian) distribution; Same covariance matrix across classes
- Quadratic Discriminant Analysis (QDA)
  - Normal (Gaussian) distribution; Different covariance matrices for each class
- Naïve Bayes
  - No assumption about a particular distribution; Xs are independent within each class

Bias-Variance trade-off

- Variance refers to the amount by which our prediction would change if we estimated it using a different training data set.
- bias refers to the error that is introduced by approximating a real-life problem, which may be extremely complicated, by a much simpler model.
- There is a trade-off when we select different methods: more flexible models have lower bias, but higher variance



LDA vs. QDA vs. Naïve Bayes

- LDA is less flexible than QDA
- LDA tends to be better if there are relatively few training observations
- QDA is recommended if the training set is very large, so that the variance is not a concern; or if the LDA assumption of common covariance matrix is clearly untenable
- Naive Bayes is a good choice in a wide range of settings. Essentially, the naive Bayes assumption introduces some bias, but reduces variance, leading to a classifier that works quite well in practice

LDA vs. Logistic regression

For a two-class problem, one can show that for LDA

$$\log\left(\frac{p_1(x)}{1-p_1(x)}\right) = \log\left(\frac{p_1(x)}{p_2(x)}\right) = c_0 + c_1x_1 + \ldots + c_px_p$$

So it has the same form as logistic regression.

The difference is in how the parameters are estimated.

- We expect LDA to outperform logistic regression when the normality assumption (approximately) holds, and
- We expect logistic regression to perform better when it does not.

Evaluate Model Performance

Evaluating the performance of classification models

- Popular criteria
  - Accuracy (misclassification) rate: % of correct classifications
  - Confusion matrix
  - Lift curve/ROC curve
- Other evaluation criteria
  - Speed and scalability
  - Interpretability
  - Robustness

Accuracy (Misclassification) rate

- Accuracy rate $= \frac{\text{Number of correct classifications}}{\text{Number of instances in dataset}}$;
- $Misclassification\ rate = 1 - Accuracy\ Rate$

		True default status		
		No	Yes	Total
Predicted default status	No	9320	128	9448
	Yes	347	205	552
	Total	9667	333	10000

Confusion Matrix

- A **confusion matrix** records the source of error:
  - **Type I error: False positives**
  - **Type II error: False negatives**

Predicted class

Actual class		Positive	Negative
	Positive	True positive	False negative
	Negative	False positive	True negative

- Suppose 950 mails are sent out
- What is the accuracy rate?

Predicted class

Actual class		Respond	Do not respond
	Respond	250	40
	Do not Respond	10	650

Confusion Matrix - Evaluation

- Below shows the performance of two classifiers. Which one is better based on accuracy?

Model 1 - Predicted class

Actual class		Respond	Do not respond
	Respond	5	5
	Do not Respond	40	950

- Accuracy = (5+950)/1000 = 95.5%
- Misclassification rate = 4.5%

Model 2 - Predicted class

Actual class		Respond	Do not respond
	Respond	10	0
	Do not Respond	90	900

- Accuracy= ?
- Misclassification rate = ?

Asymmetric costs of different types of errors

- Suppose cost of mailing to a non-responder is $1, and (net) lost revenue of not mailing to a responder is $20.
- Now from cost perspective, which classifier is better?

Model 1 - Predicted class

Actual class		Respond	Do not respond
	Respond	5	5
	Do not Respond	40	950

- Cost = 5*20 + 40*1 = $140

Model 2 - Predicted class

Actual class		Respond	Do not respond
	Respond	10	0
	Do not Respond	90	900

- Cost =

The credit card default

		True Default Status		
		No	Yes	Total
Predicted	No	9644	252	9896
Default Status	Yes	23	81	104
	Total	9667	333	10000

- What is Type I error rate? What is Type II error rate?
- As a credit card company, which type of error would it like to avoid more?
- **Sensitivity**: the proportion of all positives that are correctly identified as positives - True positive rate
- **Specificity**: the proportion of all negatives that are correctly identified as negatives – True negative rate

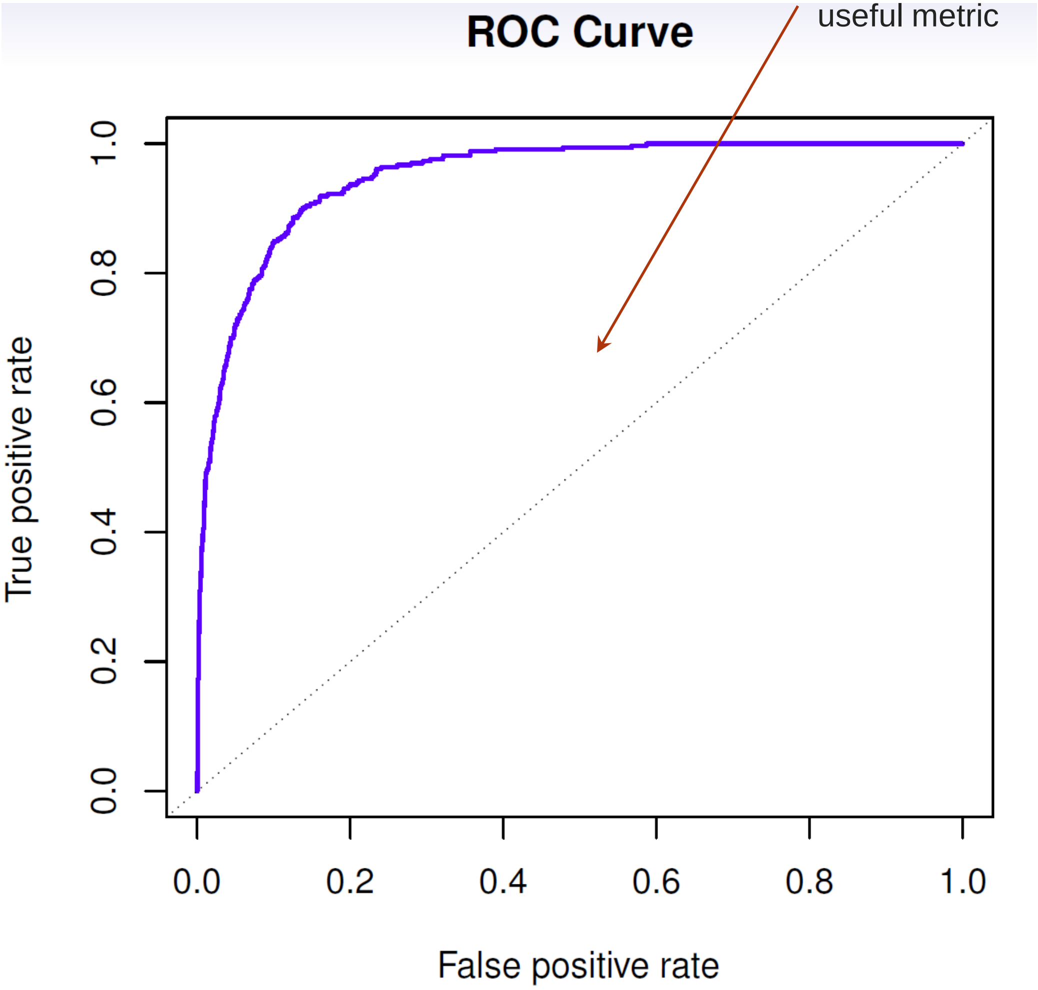
The credit card default

		True default status		
		No	Yes	Total
Predicted	No	9,432	138	9,570
default status	Yes	235	195	430
	Total	9,667	333	10,000

- We adjust the threshold probability from 0.5 to 0.2
- Sensitivity increases
- It comes at a cost of decreasing specificity and slightly increasing error rate
- There is a trade-off between sensitivity and specificity

ROC curve

- ROC curve depicts the trade-off between **Sensitivity** vs **Specificity**
- It displays two types of errors for all possible thresholds
- False positive rate: **1 - specificity**
- The overall performance is given by the area under the curve: the larger the better
- An ideal ROC curve will hug the top left corner



Area under the ROC curve (“AUC”) is a useful metric

Takeaways

- Discriminant analysis: models and assumptions
    - LDA
    - QDA
    - Naïve Bayes
- The comparison between them
- Evaluate performance of different methods
    - Accuracy rate
    - Confusion matrix
    - ROC curve