QTM 347 Machine Learning

Lecture 4: Classification

Ruoxuan Xiong

Suggested reading: ISL Chapter 4

Lecture plan

- Logistic regression
- Generative vs discriminative methods

Classification problem

- Classification is a form of supervised machine learning

- The response variable Y is categorical, as opposed to numerical for regression

- Our goal is to find a function C which takes feature(s), x, as input, and outputs a category which is the same as the true category as frequently as possible

Input x numerical or categorical	$\rightarrow$	function $\hat{C}(x)$	$\rightarrow$	Output $\hat{y}$ categorical

An example of classification problem: Image classification

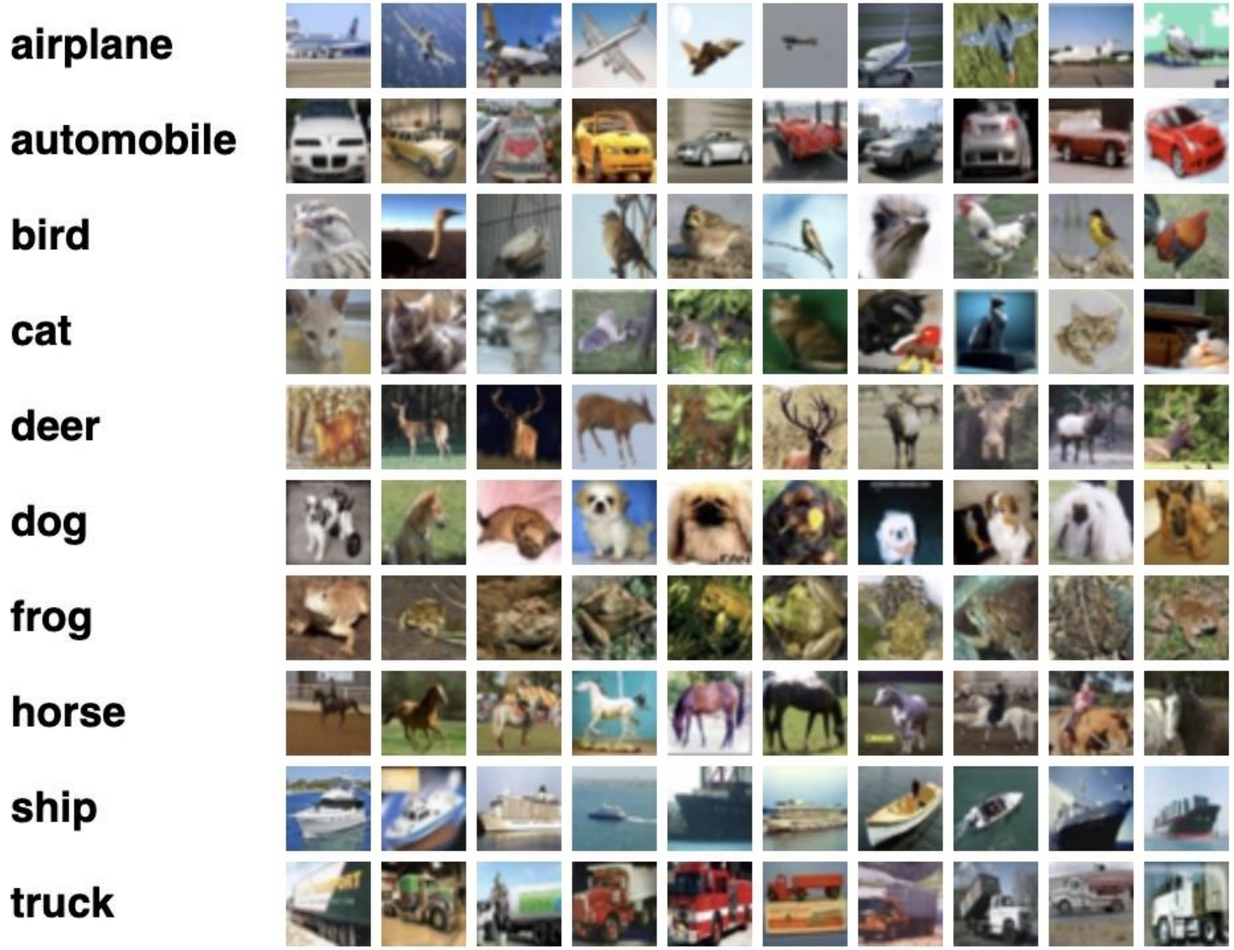
- Image classification involves assigning a label to an entire image or photograph

- For example, classifying a handwritten digit

MNIST dataset: Modified National Institute of Standards and Technology dataset
60,000 images of handwritten single digits between 0 and 9

An example of classification problem: Image classification

- Image classification involves assigning a label to an entire image or photograph
- For example, recognizing objects in photos



CIFAR-10: Canadian Institute For Advanced Research
60,000 images in 10 different classes, with 6,000 images of each class

A simpler example of classification problem

- A data set containing information on ten thousand customers
  - **default**: whether the customer defaulted on their debt
  - **student**: whether the customer is a student
  - **balance**: the average balance that the customer has remaining on their credit card after making their monthly payment
  - **income**: income of customer

- We seek to predict which customers will default on their credit card debt

```
## # A tibble: 10,000 x 4
##    default student balance income
##    <fct>   <fct>     <dbl>  <dbl>
##  1 No      No         730. 44362.
##  2 No      Yes        817. 12106.
##  3 No      No        1074. 31767.
##  4 No      No         529. 35704.
##  5 No      No         786. 38463.
##  6 No      Yes        920.  7492.
##  7 No      No         826. 24905.
##  8 No      Yes        809. 17600.
##  9 No      No        1161. 37469.
## 10 No      No           0  29275.
## # … with 9,990 more rows
```

Example: default prediction

- We seek to predict which customers will default on their credit card debt

Input		function		Output
$x =$ (student, balance, income)	→	$\hat{C}(x)$	→	$\hat{y}$ (default or not)

- What can $\hat{C}$ be?

- How can we evaluate whether $\hat{C}$ is a "good" function or not?

Example: default prediction

- We seek to predict which customers will default on their credit card debt

Input
$x = (\text{student}, \text{balance}, \text{income})$ → function
$\hat{C}(x)$ → Output
$\hat{y}$ (default or not)

- What can $\hat{C}$ be?
  - Logit model (from your regression analysis, we will have a quick review)
  - K-nearest neighbors classification
  - Linear Discriminant Analysis (LDA)
  - Quadratic Discriminant Analysis (QDA)
  - …

Example: default prediction

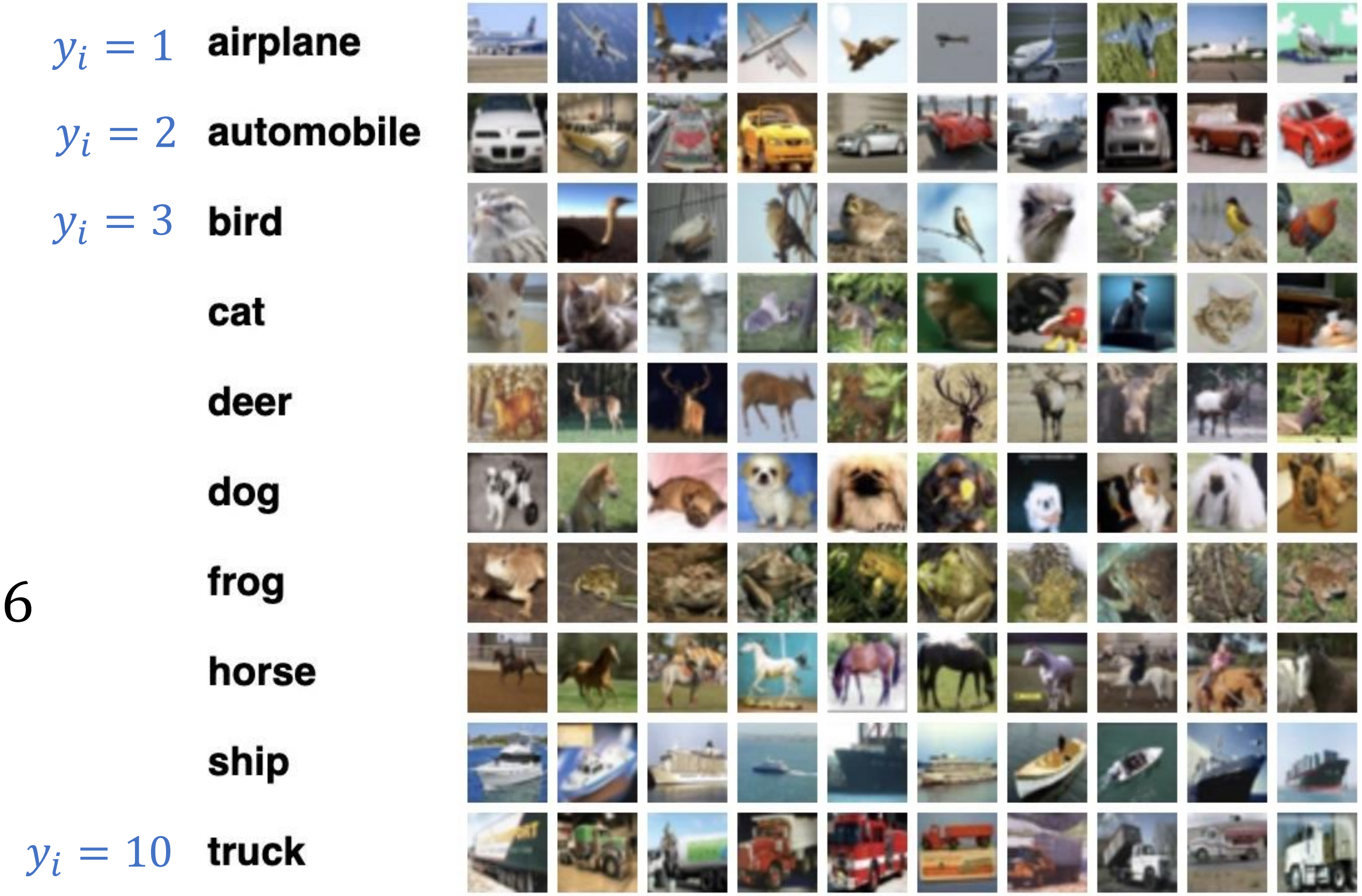
- We seek to predict which customers will default on their credit card debt

Input
$x =$ (student, balance, income) → function $\hat{C}(x)$ → Output $\hat{y}$ (default or not)

- How can we evaluate whether $\hat{C}$ is a "good" function or not?
  - We need an evaluation metric
  - Can we use MSE?

MSE is not a good metric for classification problem…

- Suppose true class $y_0 = 10$ (truck)
- We have two functions $\hat{C}$
  - First function: $\hat{y}_0 = \hat{C}_1(x_0) = 1$ (predicted as airplane)
    - Squared error: $(y_0 - \hat{y}_0)^2 = (10 - 1)^2 = 81$
  - Second function: $\hat{y}_0 = \hat{C}_2(x_0) = 6$ (predicted as dog)
    - Squared error: $(y_0 - \hat{y}_0)^2 = (10 - 6)^2 = 16$
  - Can we say first function is five times worse than second function?



CIFAR-10: Canadian Institute For Advanced Research
60,000 images in 10 different classes, with 6,000 images of each class

Evaluation metric: Classification error rate

- Classification error rate:

$$\mathrm{err}(\hat{C}) = \frac{1}{n}\sum_{i=1}^{n} 1(Y_i \neq \hat{C}(X_i))$$

- $1(\cdot)$: indicator function

$$1\left(Y_i \neq \hat{C}(X_i)\right) = \begin{cases} 1 & Y_i \neq \hat{C}(X_i) \\ 0 & Y_i = \hat{C}(X_i) \end{cases}$$

- We are essentially calculating the proportion of predicted classes that mismatch the true class

Training/test classification error rate

- Training data: the observations, $(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)$, that we use estimate C

- Training classification error rate: $\text{err}(\hat{C}) = \frac{1}{n}\sum_{i=1}^{n} 1(Y_i \neq \hat{C}(X_i))$

- Test data: the data, $(X_1', Y_1'), (X_2', Y_2'), \ldots, (X_m', Y_m')$, that are previous unseen and not used to fit C

- Test classification error rate: $\text{err}(\hat{C}) = \frac{1}{m}\sum_{i=1}^{m} 1(Y_i' \neq \hat{C}(X_i'))$

Logit model

- The logit model is (use the default prediction as an example)

$$\log\left[\frac{P(Y=1|X)}{P(Y=0|X)}\right] = \beta'\tilde{X} = \beta_0 + \beta_1 \cdot \text{student} + \beta_2 \cdot \text{balance} + \beta_3 \cdot \text{income},$$

where $\tilde{X} = (1, \text{student}, \text{balance}, \text{income})$

- $P(Y=1|X) = \frac{1}{1+e^{-(\beta_0+\beta_1\cdot\text{student}+\beta_2\cdot\text{balance}+\beta_3\cdot\text{income})}}$

- Proof: For notation simplicity, $P(Y=1|X) = p_{1X}$ and $P(Y=0|X) = p_{0X}$. Then

$$\log\left[\frac{p_{1X}}{p_{0X}}\right] = \log\left[\frac{p_{1X}}{1-p_{1X}}\right] = \beta'\tilde{X} \Rightarrow \frac{p_{1X}}{1-p_{1X}} = e^{\beta'\tilde{X}} \Rightarrow p_{1X} = e^{\beta'\tilde{X}} - p_{1X}e^{\beta'\tilde{X}}$$

$$\Rightarrow p_{1X}\left(1+e^{\beta'\tilde{X}}\right) = e^{\beta'\tilde{X}} \Rightarrow p_{1X} = \frac{e^{\beta'\tilde{X}}}{1+e^{\beta'\tilde{X}}} \Rightarrow p_{1X} = \frac{1}{1+e^{-\beta'\tilde{X}}}$$

Odd ratio

- The logit model is

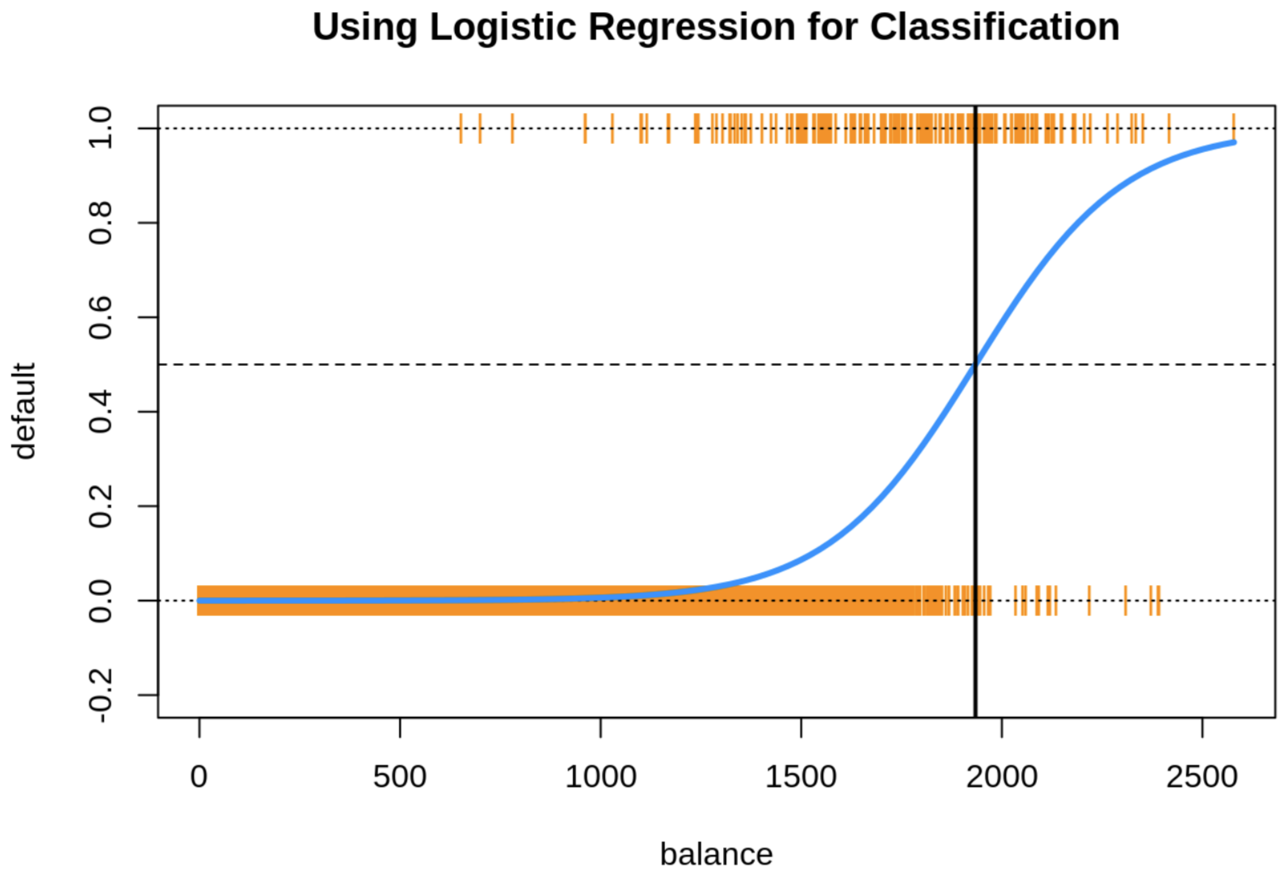
$$\log\left[\frac{P(Y=1|X)}{P(Y=0|X)}\right] = \beta'\tilde{X}$$

  - $\frac{P(Y=1|X)}{P(Y=0|X)}$: odds ratio $\in [0, \infty)$

  - $\log\left[\frac{P(Y=1|X)}{P(Y=0|X)}\right]$: log odds

How to make prediction from logit model?

- Consider a simpler logit model
  - $\log\left[\frac{P(Y=1|X)}{P(Y=0|X)}\right] = \beta_0 + \beta_1 \cdot \text{balance}$
- Fitted model
  - $\hat{P}(Y=1|X) = \frac{1}{1+e^{-(\beta_0+\beta_1 \cdot \text{balance})}}$
  - Blue curve in the figure
- Prediction
  - $\hat{C}(x) = \begin{cases} 1 & \hat{P}(Y=1|X=x) > 0.5 \\ 0 & \hat{P}(Y=1|X=x) \le 0.5 \end{cases}$



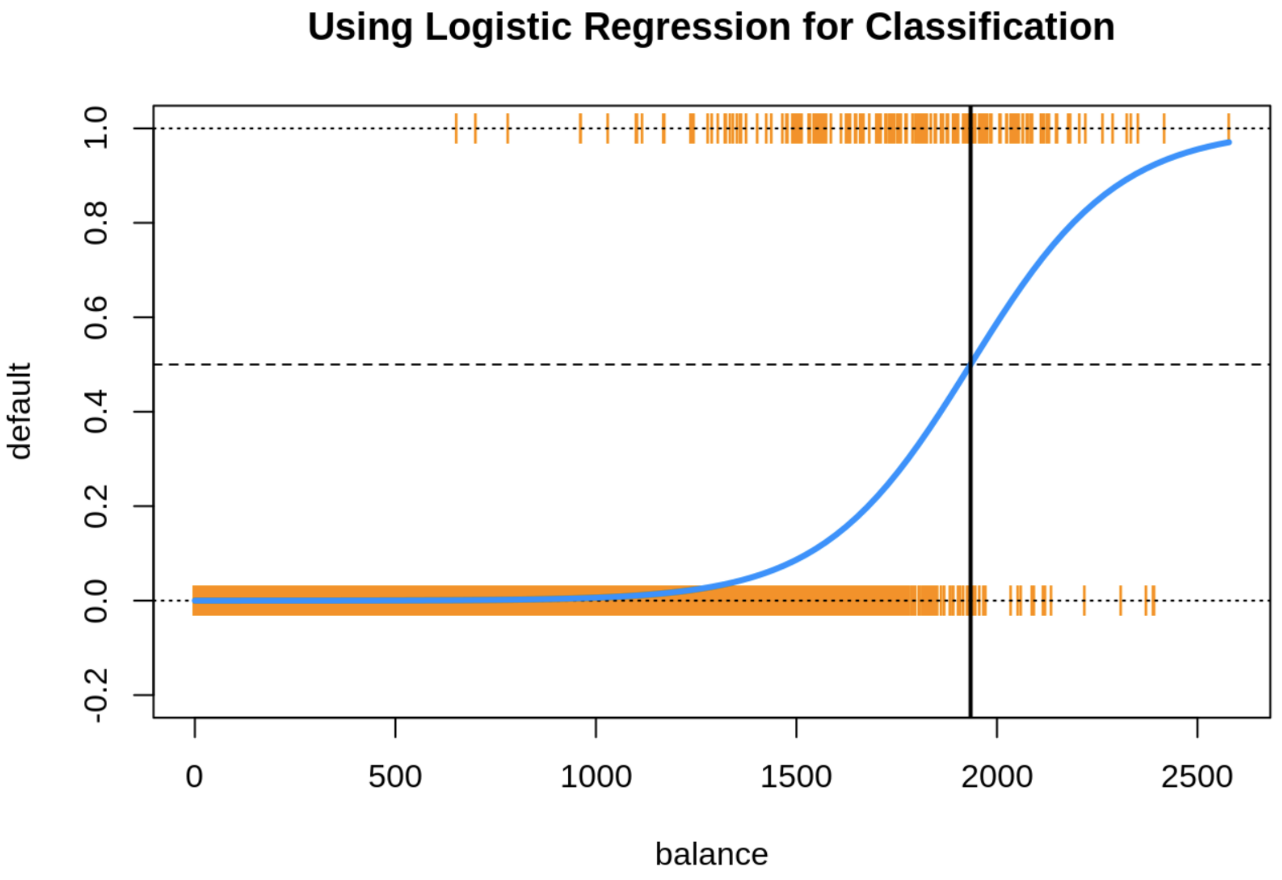
The orange | characters are the data (x_i, y_i)

How to make prediction from logit model?

- Prediction
  - $\hat{C}(x) = \begin{cases} 1 & \hat{P}(Y = 1|X = x) > 0.5 \\ 0 & \hat{P}(Y = 1|X = x) \leq 0.5 \end{cases}$

- The solid vertical black line represents the **decision boundary**, and satisfies $\hat{P}(Y = 1|X = x) = 0.5$

- In this case, $\text{balance} = 1934.22$



The orange | characters are the data (x_i, y_i)

Bayes classifier

- The rule

$$\hat{C}(x) = \begin{cases} 1 & \hat{P}(Y = 1|X = x) > 0.5 \\ 0 & \hat{P}(Y = 1|X = x) \leq 0.5 \end{cases}$$

is an example of Bayes classifier

- Bayes classifier minimizes the classification error rate
  - Proof: Suppose $\hat{P}(Y = 1|X = x) > 0.5$.
    - Output class 1: classification error rate is $1 - \hat{P}(Y = 1|X = x) < 0.5$
    - Output class 0: classification error rate is $\hat{P}(Y = 1|X = x) > 0.5$
    - ➢ Output class 1 minimizes the classification error rate

Bayes classifier

- For a general number of classes (2 or more), Bayes classifier is

$$C^B(x) = \text{argmax}_g P(Y = g|X = x)$$

  - $C^B(x)$ is the class with highest probability

  - Proof: Suppose g^* maximizes $P(Y = g|X = x)$.
    - Output class g^*: classification error rate is $1 - P(Y = g^*|X = x)$
    - Output class g': classification error rate is $1 - P(Y = g'|X = x)$
    - As $P(Y = g^*|X = x) \geq P(Y = g'|X = x)$, we have $1 - P(Y = g^*|X = x) \leq 1 - P(Y = g'|X = x)$
    - Output class g^* minimizes the classification error rate

Lecture plan

- Logistic regression

- Generative vs discriminative methods

Generative vs discriminative methods

- Generative methods
  1. Model the joint probability $p(x, y)$
  2. Assume some distribution for conditional distribution of X given $Y = k$, $P(X = x|Y = k)$
  3. Bayes theorem is applied to obtain $P(Y = k|X = x)$ and classify
  - E.g., linear discriminant analysis (LDA), quadratic discriminant analysis (QDA)

- Discriminative methods
  - Directly model $P(Y = k|X = x)$ and classify
  - E.g., logistic regression

Bayes theorem

- Given
  - Conditional distribution of X given $Y = k$: $P(X = x \mid Y = k)$
  - The **prior** probabilities for each possible class k: $P(Y = k)$

- Bayes theorem to obtain $P(Y = k \mid X = x)$

$$P(Y = k \mid X = x) = \frac{P(Y = k, X = x)}{P(X = x)}$$

$$= \frac{P(X=x|Y=k)P(Y=k)}{\sum_j P(X=x|Y=j)P\ (Y=j)}$$

Example: An iris data set

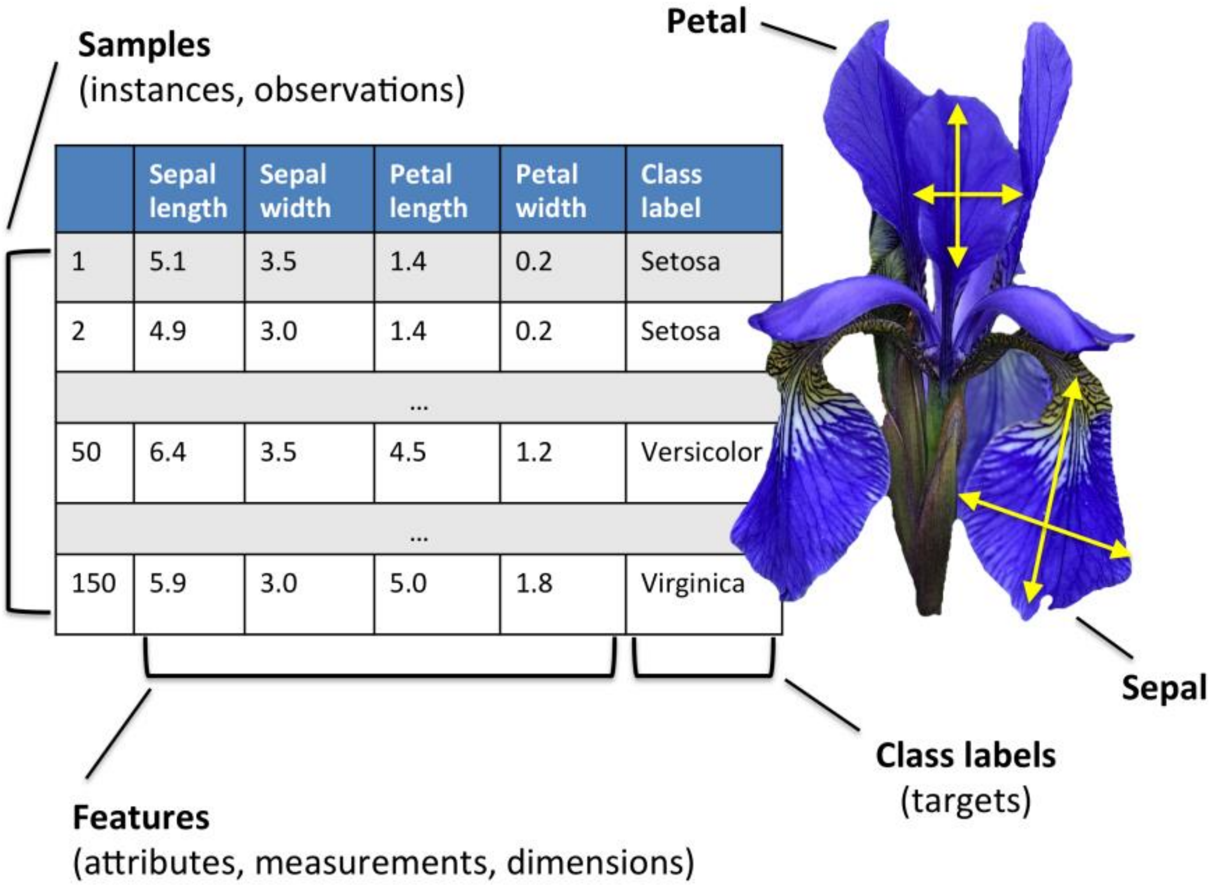
- Perhaps the best known database in the pattern recognition literature
- Predict class of iris plant
- There are three classes

Iris Versicolor **Iris Setosa** **Iris Virginica**

Sepal and petal of iris

Example: An iris data set

- **50** samples from each of three class of *Iris (versicolor, setosa, virginica)*
- Four features: sepal length, sepal width, petal length, petal width



	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
...					
50	6.4	3.5	4.5	1.2	Versicolor
...					
150	5.9	3.0	5.0	1.8	Virginica

Distribution of features