Classification

put stuff into groups

- Fisher's Iris: classifying flowers
  - three types
  - measurement: size of petal
- Analysis of political leanings of blogs:
  - mining text / NLP
- predict disease from several covariates / features
- Handwriting Digit classification
  - 16x16 image = 256 covariates / features

Binary Classification (two groups)

- people who will / will not default based on
  1. credit card balance
  2. Income

- Labels : $\{0, 1\}$

  not default, default

- binary classifier h is a function from $\mathcal{X} \longrightarrow \{0, 1\}$

  covariate space

  income & balance

- a linear classifier would be
  a straight line splitting the data

  - $H(x) = \beta_0 + \beta^T x$ such that $h(x) = I(H(x) > 0)$

    linear discriminant function

  -

- decision boundary : the line / points at which $H(x) = 0$

Bayes Risk

- Classification risk / error rate

$$R(h) = \mathbb{P}\text{robability}(\underset{\text{actual}}{Y} \neq \underset{\text{predict}}{h(X)})$$

prob of making a mistake

- training error / how many times mistake

$$\hat{R}(h) = \frac{1}{n} \sum_{i=1}^{n} I(h(x_i) \neq y_i)$$

- Theorem: the rule h that minimizes $R(h)$ is

$$h^*(x) = \begin{cases} 1 & \text{if } m(x) > 0.5 \\ 0 & \text{otherwise} \end{cases}$$

e.g if:
$P(Y=1 \mid X=x)$ 0.7
$P(Y=0 \mid X=x)$ 0.3
then you say $x \in 1$

where $m(x) = \mathbb{E}(Y \mid X=x) = \mathbb{P}(Y=1 \mid X=x)$
is the regression function

- The set $\{x \in X : m(x) = 0.5\}$ is the Baye's decision boundary

Bayes Rule

$$P(Y=y \mid X=x) = \frac{P(Y=y) \cap P(X=x)}{P(X=x)}$$

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

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

Bayes' classifier in practice

- you don't know the probabilities e.g. $P(Y=1 \mid X = x)$ whatever
  - so you need to know underlying distribution of the data or assume some sort of model

Bayes principles in k-nearest neighbours

if you don't know the priors, you can still take an average by looking at surrounding points, for each point.

Logistic Regression

Linear model

$$y = X\beta + \varepsilon$$

where $y \in \mathbb{R}$

Data-one prediction

$$P(Y_i = 1 \mid X = x_i) = p(x_i)$$

$$P(Y_i = 0 \mid X = x_i) = 1 - p(x_i)$$

we want relation between $p(x_i)$ and x_i

Logit transform

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$$

maps $[0,1]$ to $(-\infty, \infty)$

Logistic regression is linear model followed by Logit transform

Decision boundary is linear in X