Contents

Classification

- Suppose we have a collection of objects and each has an unknown **label** associated with it, e.g. like marmite or doesn't
- For a subset of the objects we observe the label plus some other properties e.g. location, nationality (features, explanatory variables, independent variables)
  - This is our **training data**
- We are willing to make a number of assumptions, our **model**
- We now want to build a **classifier** that predicts the label of a new object drawn from the collection

Examples:

- Based on the test within an email, predict whether is it spam or not
- Given the contents of my shopping basket, predict whether I am a vegetarian or not
- Given where I like in Dublin, predict which political party I'll vote for

Logistic Regression

- Label Y only takes values 0 or 1
- Real-valued vector $\vec{X}$ or m observed features $X^{(1)}, X^{(2)}, \ldots, X^{(m)}$
- In **Logistic regression** our statistical model is that:

$$P(Y = 1 \mid \Theta = \vec{\theta}, \vec{X} = \vec{x}) = \frac{1}{1 + \exp(-z)} \text{ with } z = \sum_{i=1}^{m} \theta^{(i)} x^{(i)}$$

$$P(Y = 0 \mid \Theta = \vec{\theta}, \vec{X} = \vec{x}) = 1 - P(Y = 1 \mid \Theta = \vec{\theta}, \vec{X} = \vec{x}) = \frac{\exp(-z)}{1 + \exp(-z)}$$

- Model has m parameters $\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(m)}$
  - We rather these together into a vector $\vec{\theta}$
- Will streamline notation for $P(Y = 1 \mid \Theta = \vec{\theta}, \vec{X} = \vec{x})$ to $P(Y = 1 \mid \vec{\theta}, \vec{x})$
- $P(Y = 1 \mid \vec{\theta}, \vec{x})$ changes smoothly with $\vec{x}$
- Want to try to learn to predict when $Y = 1$ and $Y = 0$ given a value of $\vec{x}$

Linear Separability

- Can also plot $P(Y = 1 \mid \vec{\Theta}, \vec{x})$ against $\vec{x}$ rather than z
- In general $\sum_{i=1}^{m} \theta^{(i)} x^{(i)} = 1$ is called a **linear** equation
  - It defines a place in m-dimensions
- Logistic regression thresholds z and predicts $Y = 1$ when $z > 0$ and $Y = 0$ when $z < 0$
- So we can think of logistic regression as trying to fit a plane that separates the $Y = 1$ data from the $Y = 0$ data
- We call such data "linearly separable"
  - Not all data is linearly separable

Parameter Estimation

- Training data is RV D. Consists of n observations $d = \{(\vec{x}_1, y), \ldots, (\vec{x}_n, y_n)\}$

Recall Bayes Rule

$$P(\Theta = \vec{\theta} \mid D = d) = \frac{P(D = d \mid \Theta = \vec{\theta} P(\Theta = \vec{\theta})}{P(D = d)}$$

- **Maximum A posteriori** (MAP) estimate of $\vec{\theta}$ is value that maximises $P(\Theta = \vec{\theta} \mid D = d)$
- Likelihood is
  $$P(D = d \mid \Theta = \vec{\theta}) = \prod_{k=1}^{n} P(Y = y_k \mid \vec{\theta}, \vec{x}_k) = \prod_{k=1}^{n} (\frac{1}{1 + \exp(-z_k)})^{y_k} (\frac{\exp(-z_k)}{1 + \exp(-z_k)})^{1 - y_k}$$
  with $z_k = \sum_{i=1}^{m} \theta^{(i)} x_k^{(i)}$
- Prior $P(\Theta = \vec{\theta})$
  - If $\vec{\theta}$ discrete valued then we can use any prior we like
  - But usually allow $\vec{\theta}$ to be continuous valued in Logistic regression
- For now let's consider **Maximum Likelihood** estimate of $\vec{\theta}$, the value which maximises $P(D \mid \vec{\theta})$

Maximum Likelihood Estimate

- Maximum Likelihood estimate is the value of $\vec{\theta}$ which maximises $P(D \mid \vec{\theta})$
- Maximising $\log P(D = d \mid \Theta = \vec{\theta}$ is the same as maximising $P(D = d \mid \Theta = \vec{\theta})$
- $\log P(D = d \mid \Theta = \vec{\theta})$ is referred to as the **log-likelihood**
- $\log P(D = d \mid \Theta = \vec{\theta}) = \log(p_1 \times (1 - p_2)) = \log p_1 + \log(1 - p_2)$ with $p_1 = \frac{1}{1+\exp(-\theta)}$, $p_2 = \frac{1}{1+\exp(+\theta)}$
- Log-likelihood maximised by selected $\theta = +\infty$. What does this mean>
- $p_1 = \frac{1}{1+\exp(-\theta)} = 1, p_2 = \frac{1}{1+\exp(+\theta)} = 0$
- So our prediction is

$$P(Y = 1 \mid \Theta = \infty, \vec{X} = \vec{x}) = \frac{1}{1 + \exp(-x)}, z = \theta^{(1)} x^{(1)} = \begin{cases} 1 & x^{(1)} = -1 \\ 0 & x^{(1)} = 0 \end{cases}$$

$$P(Y = 0 \mid \Theta = \infty, \vec{X} = \vec{x}) = 1 - P(Y = 1 \mid \Theta = \infty \vec{X} = \vec{x}) = \frac{\exp(-z)}{1 + \exp(-z)}$$

- Recall training data is $(x_1 = 1, y_1 = 1)$ and $(x_2 = -1, y_2 = 0)$

When $\vec{\theta}$ has many elements

Log-likelihood is concase, has a single maximum