Contents

Regression vs Classification **1**

Linear Models **1**

Parameter Estimation **3**

Maximum Likelihood Estimation 3

Bayesian Estimation **3**

MAP Estimation 3

MAP vs Maximum Likelihood Estimation **4**

Regression vs Classification

In classification we have

- Vector $\vec{x}$ of m observed features $x^{(1)}, x^{(2)}, \ldots, m^{(m)}$, e.g. blood pressure, age, cholestrol
- Label Y we are tyring to predict, a finite set of possible values, e.g. heart condition
- Model: Assumed statistic relationship between features $\vec{x}$ are label Y

Alternatively Y is a continuous valued random variable, so may be real-valued and:

- Prediction is now usually referred to as **regression** (rather than classification)
- Quantity Y is often referred to as the **output** or **dependent variable** (rather than the label)

Linear Models

Linear models are very popular for regression as easy to work with

- Assume a linear relationship between observed features vector $\vec{x}$ and depdendent variables Y

$$Y = \sum_{i=1}^{m} \Theta^{(i)} x^{(i)} + M$$

where $\vec{\Theta}$ is a vector of unknown (random) parameters and M is random "noise"

- Vector $\vec{\Theta}$ is unknown and we want to estimate it
- To estimate $\vec{\Theta}$ we need some **training data**, i.e.
  - A set of observations consisting of pairs of values $(\vec{x}_1, Y_1), (\vec{x}_2, Y_2), \ldots, (\vec{x}_n, Y_n)$
  - We assume that $Y_1 \sum_{i=1}^{m} \Theta^{(i)} x_1^{(i)} + M_1$ where M_1 is noise, $Y_2 = \sum_{i=1}^{m} \Theta^{(i)} x_2^{(i)} + M_2$, etc.
  - Observe that $\vec{\Theta}$ is the same for every pair of observations but that noise M_1, M_2, etc. varies
- Plus the prior distributions of Θ and M. For now we will assume:
  - M is Gaussian with mean 0 and variance 1, $\Theta^{(i)} \sim N(0, \lambda)$ (recall Y is a Normal random variables $Y \sim N(\mu, \sigma^2)$ when it has PDF $f_Y(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$)

Example: generalised linear model

- Suppose have single input x and output is

$$Y = \Theta^{(1)} x + \Theta^{(2)} x^2 + \cdots + \Theta^{(m)} s^m + M, M \sim N(0,1), \Theta^{(i)} \sim N(0, \lambda)$$

- Define feature vector $\vec{Z}$ with $z^{(1)} = x, z^{(2)} = x^2, \ldots, z^{(m)} = x^m$
- Using this vector the model is

$$Y = \sum_{i=1}^{m} \Theta^{(i)} z^{(i)} + M$$

Although model is nonlinear in x it is linear in $\vec{z}$. These new $\vec{z}$ can be computed given input x, so its known.

There's another way to write linear model in terms of PDFs.

- Previous used $Y = \vec{\Theta} x + M, M \sim N(0,1), \Theta^{(i)} \sim N(0, \lambda)$
- Given $\vec{\Theta} = \vec{\theta}$, then $Y - \sum_{i=1}^{m} \theta^{(i)} x^{(i)} = M \sim N(0,1)$ i.e.

$$f_{Y|X,\vec{\Theta}}(y \mid x, \vec{\theta}) = \frac{1}{\sqrt{2\pi}} \exp(-(y - \sum_{i=1}^{m} \theta^{(i)} x^{(i)})^2/2)$$

- Note that we have to use PDF rather than PMF since Y is a continuous RV
- Model also assumes $\Theta^{(i)} \sim N(0, \lambda)$ i.e.

$$f_{\Theta^{(i)}}(\theta) \propto \exp(-\theta^2/2\lambda)$$

- $f_{Y|X,\vec{\Theta}}(y \mid x, \vec{\theta})$ and $f_{\Theta^{(i)}}(\theta^{(i)})$ fully describe the linear model

Parameter Estimation

Recall Bayes Rule for PDFs

$$f_{\Theta|D}(\vec{\Theta} \mid d) = \frac{f_{D|\Theta}(d \mid \vec{\theta} f_{\Theta}(\vec{\theta})}{f_D(d)}$$

- Likelihood: $f_{D|\Theta}(d \mid \vec{\theta})$

Maximum Likelihood Estimation

Select the value $\vec{\theta}$ which maximises likelihood $f_{D|\Theta}(d \mid \vec{\theta})$

- $Y = \sum_{i=1}^{m} \Theta^{(i)} x^{(i)} + M, M \sim N(0,1), \Theta^{(i)} \sim N(0, \lambda)$
- Conditioned on $\vec{\Theta} = \vec{\theta}$ we have

  $$f_{D|\Theta}(d \mid \vec{\theta}) \propto L(\theta) = \exp(-\sum_{j=1}^{n}(y_j - \sum_{i=1}^{m} \theta^{(i)} x_j^{(i)})^2/2)$$

  dropping the normalising constant as it doesn't matter here
- Take log (giving the "log-likelihood"):

  $$\log f_{D|\Theta}(d \mid \vec{\theta}) \propto \log L(\theta) = -\frac{1}{2}\sum_{j=1}^{n}(y_j - \sum_{i=1}^{m} \theta^{(i)} x_j^{(i)})^2$$

- We want to select $\vec{\theta}$ to maximise $\log L(\vec{\theta})$ i.e. the minimise $\sum_{j=1}^{n}(y_j - \sum_{i=1}^{m} \theta^{(i)} x_j^{(i)})^2$
- Called the "least squares" estimate, for obvious reasons

Bayesian Estimation

- Estimate the posterier $f_{\Theta|D}(\theta \mid d)$, rather than the likelihood $f_{D|\theta}(d \mid \theta)$
- A *distribution* rather than just a singple value

MAP Estimation

- Maximum a porteriori (MAP) estimation
- Selection θ that maximises posterior $f_{\Theta|D}(\theta \mid d)$ (back to a single value rather than a distribution)
- Runs into trouble is distribution has > 1 peak

- Map estimate:

$$\theta = \frac{\sum_{j=1}^{n} y_j x_j}{\frac{1}{\lambda} + \sum_{j=1}^{n} x_j^2}$$

- May estimate depends on our choice of λ
  - Remember that this value reflects our prior belief of the distribution of parameter Θ, $f_\Theta(\theta) \propto \exp(-\theta^2/2\lambda)$
- When $\lambda = 0$ then we are saying that we are *certain* Θ is 0 ($\theta = \frac{\sum_{j=1}^{n} y_j x_j}{\frac{1}{\lambda}+\sum_{j=1}^{n} x_j^2} \to 0$ as $\lambda \to 0$)
- When λ is very large we are saying that we know very little about the value of Θ prior to making the observations
- MAP estimate is then close to the maximum likelihood estimate

MAP vs Maximum Likelihood Estimation

Difference between MAP and ML really kicks in when we only have a small number of observations, yet still need to make a prediction. Our prior beliefs are then especially important.

- But as number N of observations grows, impact of prior on posterior tends to decline
- Remember two interpretations of probability, as frequency and belief respectively
- Important when we need to make a decision with limited data