Overview

- Quick Probability Refresh
- Probabilistic Interpretation of Linear Regression
- Probabilistic Interpretation of Logistic Regression
- Probabilistic Interpretation of Regularisation

Probability Refresh

- **Sample space** S: set of possible outcomes, **random event** E: subset of S, **random variable**: maps event E to a real value.
- Can think of probability of an event E as the frequency with which it happens when an experiment is repeated many times
- **Conditional probability**:
  - Events: $P(E|F) = \frac{P(E \cap F)}{P(F)}$ when $P(F) > 0$.
  - RVs: $P(X = x|Y = y) = \frac{P(X=x \text{ and } Y=y)}{P(Y=y)}$
- **Chain rule**: $P(X = x \text{ and } Y = y) = P(X = x|Y = y)P(Y = y)$.

Probability Refresh

Consequences of chain rule:

- **Marginalisation**:
  Suppose RV Y takes values in $\{y_1, y_2, ..., y_n\}$. Then

$$\begin{aligned} P(X = x) &= P(X = x \text{ and } Y = y_1) + \cdots + P(X = x \text{ and } Y = y_n) \\ &= \sum_{i=1}^{n} P(X = x | Y = y_i) P(Y = y_i) \end{aligned}$$

- **Bayes rule**:
  $P(X = x | Y = y) = \frac{P(Y=y|X=x)P(X=x)}{P(Y=y)}$.
- **Independence**: Random variables X and Y are independent if

$$P(X = x \text{ and } Y = y) = P(X = x) P(Y = y)$$

  for all x and y, in which case $P(X = x | Y = y) = P(X = x)$.

Probability Refresh

Continuous-valued random variables:

- $P(X = x) = 0$ for continuous-valued random variables, instead we need to consider intervals e.g. $P(a \leq X \leq b)$.
- $F_Y(y) := P(Y \leq y)$ is the **cumulative distribution function** (CDF) and $P(a < Y \leq b) = F_Y(b) - F_Y(a)$.
- For a continuous-valued random variable Y there exists a **probability density function** $f_Y(y) \geq 0$ such that:

  $$F_Y(y) = \int_{-\infty}^{y} f_Y(t)dt$$

  and so

  $$P(a < Y \leq b) = \int_{-\infty}^{b} f_Y(t)dt - \int_{-\infty}^{a} f_Y(t)dt = \int_{a}^{b} f_Y(t)dt$$

- The probability density function $f(y)$ for random variable Y is <u>not</u> a probability e.g. it can take values greater than 1. Its the <u>area</u> under the PDF that is the probability $P(a < Y \leq b)$
- $\int_{-\infty}^{\infty} f(y)dy = 1$ (since $\int_{-\infty}^{\infty} f(y)dy = F_Y(\infty) = P(Y \leq \infty) = 1$)

Probability Refresh

- $F_{XY}(x, y) = P(X \leq x \text{ and } Y \leq y)$ is the cumulative distribution function for X and Y. It is well-defined for both continuous and discrete valued RVs
- When X and Y are continuous-valued random variables there exists a probability density function (PDF) $f_{XY}(x, y) \geq 0$ such that:
  - $F_{XY}(x, y) = \int_{\infty}^{x} \int_{-\infty}^{y} f_{XY}(u, v) du\ dv$
- Define conditional PDF:

$$f_{X|Y}(x|y) = \frac{f_{XY}(x, y)}{f_Y(y)}$$

- Then chain rule also holds for PDFs:

$$f_{XY}(x, y) = f_{X|Y}(x|y) f_Y(y) = f_{Y|X}(y|x) f_X(x)$$

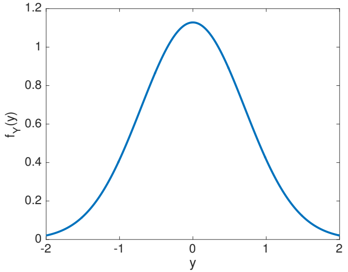
- So marginalisation, Bayes rule and independence carry over to PDFs similarly to discrete-valued RVs

Probability Refresh

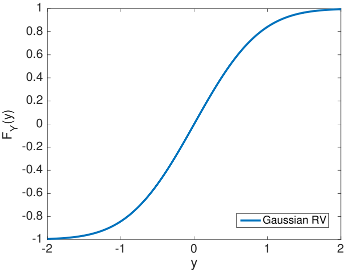
Y is a **Normal** or **Gaussian** random variable $Y \sim N(\mu, \sigma^2)$ when it has PDF:

$$f_Y(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$$

PDF



CDF



$\mu = 0,\ \sigma = 1$

- $E[Y] = \mu,\ Var(Y) = \sigma^2$
- Symmetric about μ and defined for all real-valued x

Probabilistic Interpretation: Linear Regression

- Assume output y is generated by:

$$Y = \theta^T x + M = h_\theta(x) + M$$

where $h_\theta(x) = \theta^T x$ and M is Gaussian noise with mean 0 and variance 1. As usual, we use capitals for random variables.

- So training data d is:

$$\{(x^{(1)}, h_\theta(x^{(1)}) + M^{(1)}), (x^{(2)}, h_\theta(x^{(2)}) + M^{(2)}), \cdots, (x^{(m)}, h_\theta(x^{(m)}) + M^{(m)})\}$$

where $M^{(1)}, M^{(2)}, \ldots, M^{(m)}$ are **independent** random variables each of which is Gaussian with mean 0 and variance 1.

Probabilistic Interpretation: Linear Regression

- A Gaussian RV Z with mean μ and variance σ^2 has pdf $f_Z(z) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(Z-\mu)^2}{2\sigma^2}}$.
- So we are assuming: $f_M(m) = \frac{1}{\sqrt{2\pi}} e^{-\frac{m^2}{2}}$, $f_Y(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(y-h_\theta(x))^2}{2}}$.
- The **likelihood** $f_{D|\Theta}(d|\theta)$ of the training data d is therefore:

$$f_{D|\Theta}(d|\theta) = \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}} e^{-\frac{(y^{(i)}-h_\theta(x^{(i)}))^2}{2}}$$

- Taking logs: $\log f_{D|\Theta}(d|\theta) = \log \frac{1}{\sqrt{2\pi}} - \sum_{i=1}^{m} \frac{(y^{(i)}-h_\theta(x^{(i)}))^2}{2}$
- And the maximum likelihood estimate of θ maximises

$$\max_\theta - \sum_{i=1}^{m} (y^{(i)} - h_\theta(x^{(i)}))^2$$

i.e. minimises

$$\min_\theta \sum_{i=1}^{m} (y^{(i)} - h_\theta(x^{(i)}))^2$$

Probabilistic Interpretation: Who Cares ?

- Since probability is about reasoning under uncertainty it would be v odd indeed if our machine learning algorithms did not make good sense from a probability/statistics point of view.
- Casting an ML approach within a statistical framework clarifies the assumptions that have been made (perhaps implicitly). E.g. in linear regression:
  - Noise is additive $Y = \theta^T x + M$
  - Noise on each observation is independent and identically distributed
  - Noise is Gaussian – it is this which leads directly to the use of a square loss $(y - h_\theta(x))^2$. Changing the noise model would lead to a different loss function.
- We can leverage the wealth of results and approaches from probability/statistics, and perhaps gain new insights. E.g. in linear regression:
  - Without regularisation, our estimate of θ is the maximum likelihood estimate. Would a MAP estimate be more/less useful ?

Probabilistic Interpretation: Logistic Regression

- Assume

$$P(Y=y|\theta,x)=\frac{1}{1+e^{-y\theta^T x}}$$

  and recall $y=1$ or $y=-1$ only.
- The **likelihood** $f_{D|\Theta}(d|\theta)$ of the training data d is therefore:

$$f_{D|\Theta}(d|\theta)=\prod_{i=1}^{m}\frac{1}{1+e^{-y\theta^T x}}$$

- Taking logs:

$$\log f_{D|\Theta}(d|\theta)=\sum_{i=1}^{m}\log\frac{1}{1+e^{-y\theta^T x}}$$

- And the maximum likelihood estimate of θ minimises:

$$-\sum_{i=1}^{m}\log\frac{1}{1+e^{-y\theta^T x}}=\sum_{i=1}^{m}\log(1+e^{-y\theta^T x})$$

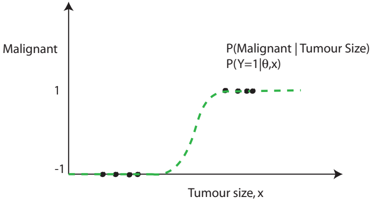
  since $-\log(z)=\log(1/z)$.

Probabilistic Interpretation: Logistic Regression

- The probabilistic formulation of logistic regression provides us with a new insight:

$$P(Y = y|\theta, x) = \frac{1}{1 + e^{-y\theta^T x}}$$

- So in addition to prediction $h_\theta(x) = sign(\theta^T x)$ we also have an estmate of our confidence in the prediction, namely $\frac{1}{1+e^{-y\theta^T x}}$.



- When $\frac{1}{1+e^{-y\theta^T x}}$ is close to 1, then we are confident in our prediction but when $\frac{1}{1+e^{-y\theta^T x}}$ is small then we are less confident.

Probabilistic Interpretation: Regularisation

- Recall Bayes Rule

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

posterior *likelihood* *prior*

- Likelihood. Probability of seeing the data d given model with parameter $\Theta = \vec{\theta}$
- Prior. Before seeing any data what is our belief about the model i.e. what is probability of parameter values Θ.
- Posterior. After seeing the data, what is our belief about probability of parameter values Θ now that we have seen the data.
- **Maximum A Posteriori** (MAP) estimate of $\vec{\theta}$ is value that maximises $P(\Theta = \vec{\theta}|D = d)$

Probabilistic Interpretation: Regularisation

- Maximum Likeihood estimation: select value of θ that maximises $P(D = d|\Theta = \theta)$
- Maximum a posteriori (MAP) estimation: select θ that maximises $P(\Theta = \theta|D = d)$.
- Taking logs in Bayes Rule:

$$\log P(\Theta = \theta|D = d) = \log P(D = d|\Theta = \theta) + \log P(\Theta = \theta) \\ - \log P(D = d)$$

Can drop the $\log P(D = d)$ term since d is fixed, so we select θ to maximise:

$$\underbrace{\log P(D = d|\Theta = \theta)}_{log-likelihood} + \underbrace{\log P(\Theta = \theta)}_{log-prior}$$

or for continuous-valued RVs:

$$\underbrace{\log f_{D|\Theta}(D = d|\Theta = \theta)}_{log-likelihood} + \underbrace{\log f_{\Theta}(\Theta = \theta)}_{log-prior}$$

Probabilistic Interpretation: Regularisation

Ridge regression variant of linear regression:

- $Y = \Theta x + M$, $M \sim N(0,1)$ as before.
- $\Theta_j, \sim N(0,\sigma^2)$ (this is our prior on θ_j), $j = 1,\dots,n$
- log-likelihood: $-\sum_{i=1}^{m}(y^{(i)} - \theta^T x^{(i)})^2$
- log-prior: $-\theta_j^2/\sigma^2$
- So MAP estimate selects θ to maximise:

$$-\sum_{i=1}^{m}(y^{(i)} - \theta^T x^{(i)})^2 - \sum_{j=1}^{n}\theta_j^2/\sigma^2$$

i.e. to minimise:

$$\sum_{i=1}^{m}(y^{(i)} - \theta^T x^{(i)})^2 + \sum_{j=1}^{n}\theta_j^2/\sigma^2$$