Lecture 5 - *Probability Revision*

OBJECTIVE: Revise the fundamental concepts of probability, including marginalization, conditioning, Bayes rule and expectation.

◇ PROBABILITY

Probability theory is the formal study of the laws of chance. It is our tool for dealing with uncertainty. Notation:

- **Sample space:** is the set Ω of all outcomes of an experiment.
- **Outcome:** what we observed. We use $\omega \in \Omega$ to denote a particular outcome. *e*.g. for a die we have $\Omega = \{1, 2, 3, 4, 5, 6\}$ and ω could be any of these six numbers.
- **Event:** is a subset of Ω that is well defined (measurable). *e.g.* the event $A = \{even\}$ if $w \in \{2, 4, 6\}$

Why do we need measure?

⋆

Frequentist Perspective

Let probability be the frequency of events.

⋆

Axiomatic Perspective

The frequentist interpretation has some shortcomings when we ask ourselves questions like

- *what is the probability that David will sleep with Anne?*
- *What is the probability that the Panama Canal is longer than the Suez Canal?*

The axiomatic view is a more elegant mathematical solution. Here, a **probabilistic model** consists of the triple $(\Omega, \mathcal{F}, P)$, where Ω is the sample space, $\mathcal{F}$ is the sigma-field (collection of measurable events) and P is a function mapping $\mathcal{F}$ to the interval $[0, 1]$. That is, with each event $A \in \mathcal{F}$ we associate a probability $P(A)$.

Some outcomes are not measurable so we have to assign probabilities to $\mathcal{F}$ and not Ω. Fortunately, in this course everything will be measurable so we need no concern ourselves with measure theory. We do have to make sure the following two axioms apply:

1. $P(\emptyset) = 0 \leq p(A) \leq 1 = P(\Omega)$
2. For **disjoint sets** A_n, $n \geq 1$, we have

$$P\left(\sum_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P(A_n)$$

★

If the sets overlap:

★

$$P(A + B) = P(A) + P(B) - P(AB)$$

If the events A and B are **independent**, we have $P(AB) = P(A)P(B)$.

⋆ Let $P(HIV) = 1/500$ be the probability of contracting HIV by having unprotected sex. If one has unprotected sex twice, the probability of contracting HIV becomes:

What if we have unprotected sex 500 times?

Conditional Probability

$$P(A|B) \triangleq \frac{P(AB)}{P(B)}$$

where $P(A|B)$ is the **conditional probability** of A given that B occurs, $P(B)$ is the **marginal probability** of B and $P(AB)$ is the **joint probability** of A and B. In general, we obtain a **chain rule**

$$P(A_{1:n}) = P(A_n|A_{1:n-1})P(A_{n-1}|A_{1:n-2})\dots P(A_2|A_1)P(A_1)$$

⋆ Assume we have an urn with 3 red balls and 1 blue ball: $U = \{r, r, r, b\}$. What is the probability of drawing (without replacement) 2 red balls in the first 2 tries?

Marginalisation

Let the sets $B_{1:n}$ be disjoint and $\bigcup_{i=1}^{n} B_i = \Omega$. Then

$$P(A) = \sum_{i=1}^{n} P(A, B_i)$$

⋆ Proof:

⋆ What is the probability that the second ball drawn from our urn will be red?

Bayes Rule

Bayes rule allows us to reverse probabilities:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Combinining this with marginalisation, we obtain a powerful tool for statistical modelling:

$$P(model_i|data) = \frac{P(data|model_i)P(model_i)}{\sum_{j=1}^{M} P(data|model_j)P(model_j)}$$

That is, if we have **prior** probabilities for each model and generative data models, we can compute how likely each model is **a posteriori** (in light of our prior knowledge and the evidence brought in by the data).

Discrete random variables

Let E be a discrete set, e.g. $E = \{0, 1\}$. A **discrete random variable** (r.v.) is a map from Ω to E:

$$X(w) : \Omega \mapsto E$$

such that for all $x \in E$ we have $\{w | X(w) \leq x\} \in \mathcal{F}$. Since $\mathcal{F}$ denotes the measurable sets, this condition simply says that we can compute (measure) the probability $P(X = x)$.

⋆ Assume we are throwing a die and are interested in the events $E = \{even, odd\}$. Here $\Omega = \{1, 2, 3, 4, 5, 6\}$. The r.v. takes the value $X(w) = even$ if $w \in \{2, 4, 6\}$ and $X(w) = odd$ if $w \in \{1, 3, 5\}$. We describe this r.v. with a **probability distribution** $p(x_i) = P(X = x_i) = \frac{1}{2}, i = 1, \ldots, 2$

The **cumulative distribution function** is defined as $F(x) = P(X \leq x)$ and would for this example be:

⋆

Bernoulli Random Variables

Let $E = \{0, 1\}$, $P(X = 1) = \lambda$, and $P(X = 0) = 1 - \lambda$. We now introduce the *set indicator variable*. (This is a very useful notation.)

$$\mathbb{I}_A(w) = \begin{cases} 1 \;\; if & w \in A; \\ 0 \;\; otherwise. \end{cases}$$

Using this convention, the probability distribution of a **Bernoulli** random variable reads:

$$p(x) = \lambda^{\mathbb{I}_{\{1\}}(x)}(1 - \lambda)^{\mathbb{I}_{\{0\}}(x)}.$$

Expectation of Discrete Random Variables

The expectation of a discrete random variable X is

$$\mathbb{E}[X] = \sum_E x_i p(x_i)$$

The expectation operator is linear, so $\mathbb{E}(ax_1 + bx_2) = a\mathbb{E}(x_1) + b\mathbb{E}(x_2)$. In general, the expectation of a function $f(X)$ is

$$\mathbb{E}[f(X)] = \sum_E f(x_i)\, p(x_i)$$

Mean: $\mu \triangleq \mathbb{E}(X)$

Variance: $\sigma^2 \triangleq \mathbb{E}[(X - \mu)^2]$

⋆ For the set indicator variable $\mathbb{I}_A(\omega)$,

$$\mathbb{E}[\mathbb{I}_A(\omega)] =$$

Continuous Random Variables

A continuous r.v. is a map to a continuous space, $X(w) : \Omega \mapsto \mathbb{R}$, under the usual measurability conditions. The **cumulative distribution function** $F(x)$ (cdf) is defined by

$$F(x) \triangleq \int_{-\infty}^{x} p(y)\, dy = P(X \leq x)$$

where $p(x)$ denotes the **probability density function** (pdf). For an infinitesimal measure dx in the real line, distributions F and densities p are related as follows:

$$F(dx) = p(x)dx = P(X \in dx).$$

⋆

Univariate Gaussian Distribution

The pdf of a Gaussian distribution is given by

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}.$$

⋆

Our short notation for Gaussian variables is $X \sim \mathcal{N}(\mu, \sigma^2)$.

Univariate Uniform Distribution

A random variable X with a uniform distribution between 0 to 1 is written as $X \sim \mathcal{U}_{[0,1]}(x)$

$\star$

Multivariate Distributions

Let $f(u,v)$ be a pdf in 2-D. The cdf is defined by

$$F(x,y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(u,v)\, du\, dv = P(X \leq x, Y \leq y).$$

1 Bivariate Uniform Distribution

$$X \sim \mathcal{U}_{[0,1]^2}(x)$$

$\star$

Multivariate Gaussian Distribution

Let $x \in \mathbb{R}^n$. The pdf of an n-dimensional Gaussian is given by

$$p(x) = \frac{1}{2\pi^{n/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}$$

where

$$\mu = \begin{pmatrix} \mu_1 \\ : \\ \mu_n \end{pmatrix} = \begin{pmatrix} \mathbb{E}(x_1) \\ : \\ \mathbb{E}(x_n) \end{pmatrix}$$

and

$$\Sigma = \begin{pmatrix} \sigma_{11} \cdots \sigma_{1n} \\ \cdots \\ \sigma_{n1} \cdots \sigma_{nn} \end{pmatrix} = \mathbb{E}[(X-\mu)(X-\mu)^T]$$

with $\sigma_{ij} = \mathbb{E}[X_i - \mu_i)(X_j - \mu_j)^T]$.

We can interpret each component of x, for example, as a feature of an image such as colour or texture. The term $\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)$ is called the **Mahalanobis distance**. Conceptually, it measures the distance between x and μ.

⋆ What is $\int \cdots \int e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}\, dx$?

Linear Operations

Let $A \in \mathbb{R}^{k\times n}$, $b \in \mathbb{R}^k$ be given matrices, and $X \in \mathbb{R}^n$ be a random variable with mean $\mathbb{E}(X) = \mu_x \in \mathbb{R}^n$ and covariance $cov(X) = \Sigma_X \in \mathbb{R}^{n\times n}$. We define a new random variable

$$Y = AX + b$$

If $X \sim N(\mu_x, \Sigma_x)$, then $Y \sim N(\mu_y, \Sigma_y)$ where

⋆

$$\mu_y = \mathbb{E}(Y) =$$

$$\Sigma_y =$$

Finally, we define the **cross-covariance** as

$$\Sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)'].$$

X and Y are **uncorrelated** if $\Sigma_{XY} = 0$. So,

$$\Sigma = \begin{pmatrix} \Sigma_{XX} & 0 \\ 0 & \Sigma_{YY} \end{pmatrix}.$$

Lecture 6 - *Linear Supervised Learning*

OBJECTIVE: Linear regression is a supervised learning task. It is of great interest because:

- Many real processes can be approximated with linear models.
- Linear regression appears as part of larger problems.
- It can be solved analytically.
- It illustrates many of the ideas in machine learning.

Given the data $\{x_{1:n}, y_{1:n}\}$, with $x_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$, we want to fit a hyper-plane that maps x to y.

★

Mathematically, the linear model is expressed as follows:

$$\widehat{y}_i = \theta_0 + \sum_{j=1}^{d} x_{ij}\theta_j$$

We let $x_{i,0} = 1$ to obtain

$$\widehat{y}_i = \sum_{j=0}^{d} x_{ij}\theta_j$$

In matrix form, this expression is

$$\widehat{Y} = X\theta$$

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_{10} & \cdots & x_{1d} \\ \vdots & \vdots & \vdots \\ x_{n0} & \cdots & x_{nd} \end{bmatrix} \begin{bmatrix} \theta_0 \\ \vdots \\ \theta_d \end{bmatrix}$$

If we have several outputs $y_i \in \mathbb{R}^c$, our linear regression expression becomes:

★

We will present several approaches for computing θ.

◇ OPTIMIZATION APPROACH

Our aim is to mininimise the quadratic cost between the output labels and the model predictions

$$C(\theta) = (Y - X\theta)^T (Y - X\theta)$$

⋆

We will need the following result from matrix differentiation: $\frac{\partial A}{\partial \theta} = A^T$.

⋆

$$\frac{\partial C}{\partial \theta} =$$

These are the **normal equations**. The solution (estimate) is:

$$\widehat{\theta} =$$

The corresponding predictions are

$$\widehat{Y} = HY =$$

where H is the “hat” matrix.

◇ GEOMETRIC APPROACH

⋆

$$X^T(Y - \widehat{Y}) =$$

Maximum Likelihood

⋆

If our errors are Gaussian distributed, we can use the model

$$Y = X\theta + \mathcal{N}(0, \sigma^2 I)$$

Note that the mean of Y is $X\theta$ and that its variance is $\sigma^2 I$. So we can equivalently write this expression using the probability density of Y **given** X, θ and σ:

$$p(Y|X,\theta,\sigma) = \left(2\pi\sigma^2\right)^{-n/2} e^{-\frac{1}{2\sigma^2}(Y-X\theta)^T(Y-X\theta)}$$

The maximum likelihood (ML) estimate of θ is obtained by taking the derivative of the log-likelihood, $\log p(Y|X,\theta,\sigma)$. The idea of maximum likelihood learning is to maximise the likelihood of seeing some data Y by modifying the parameters (θ, σ).

The ML estimate of θ is:

⋆

Proceeding in the same way, the ML estimate of σ is:

⋆

Lecture 7 - *Ridge Regression*

OBJECTIVE: Here we learn a cost function for linear supervised learning that is more stable than the one in the previous lecture. We also introduce the very important notion of **regularization**.

All the answers so far are of the form

$$\widehat{\theta} = (XX^T)^{-1}X^TY$$

They require the inversion of XX^T. This can lead to problems if the system of equations is poorly conditioned. A solution is to add a small element to the diagonal:

$$\widehat{\theta} = (XX^T + \delta^2 I_d)^{-1}X^TY$$

This is the ridge regression estimate. It is the solution to the following **regularised quadratic cost function**

$$C(\theta) = (Y - X\theta)^T(Y - X\theta) + \delta^2\theta^T\theta$$

⋆ Proof:

It is useful to visualise the quadratic optimisation function and the constraint region.

⋆

That is, we are solving the following **constrained optimisation** problem:

$$\min_{\theta\,:\;\theta^T\theta \le t} \left\{(Y - X\theta)^T(Y - X\theta)\right\}$$

Large values of θ are penalised. We are **shrinking** θ towards zero. This can be used to carry out **feature weighting**. **An input $x_{i,d}$ weighted by a small θ_d will have less influence on the ouptut y_i.**

Spectral View of LS and Ridge Regression

Again, let $X \in \mathbb{R}^{n \times d}$ be factored as

$$X = U\Sigma V^T = \sum_{i=1}^{d} u_i \sigma_i v_i^T,$$

where we have assumed that the rank of X is d.

⋆ The least squares prediction is:

$$\widehat{Y}_{LS} = \sum_{i=1}^{d} u_i u_i^T Y$$

⋆ Likewise, for ridge regression we have:

$$\widehat{Y}_{ridge} = \sum_{i=1}^{d} \frac{\sigma_i^2}{\sigma_i^2 + \delta^2} u_i u_i^T Y$$

The filter factor

$$f_i = \frac{\sigma_i^2}{\sigma_i^2 + \delta^2}$$

penalises small values of σ^2 (they go to zero at a faster rate).

⋆

Also, by increasing δ^2 we are penalising the weights:

⋆

Small eigenvectors tend to be wobbly. The Ridge filter factor f_i gets rid of the wobbly eigenvectors. Therefore, the predictions tend to be more stable (smooth, regularised).

The smoothness parameter δ^2 is often estimated by cross-validation or Bayesian hierarchical methods.

Minimax and cross-validation

Cross-validation is a widely used technique for choosing δ. Here's an example:

⋆