Probability summaries

1 Probability theory

- A **sample space** is a set of point, they are the possible **outcomes** of a **trial**.
- An **event** is a subset of a sample space.
- A **probability** is a map from events to real numbers such that
  1. $P(A) \geq 0$ for all events.
  2. $P(X) = 1$
  3. If $A \cap B = \emptyset$ for two events A and B then
  $$P(A \cup B) = P(A) + P(B) \tag{1}$$
- A **probability mass function** is a map from points in the sample space to real numbers such that
  1. $p(x) \geq 0$ for all $x \in X$
  2. $\sum_{x \in X} p(x) = 1$
- $P(A) = \sum_{x \in A} p(x)$
- If all the points in a sample space have the same probability then
  $$P(A) = \frac{\text{number of points in } A}{\text{number of points in } X} = \frac{\#(A)}{\#X} \tag{2}$$
  where $\#(A)$ means the number of points in A.
- The **binomial coefficient**
  $$\left(\begin{array}{c} n \\ r \end{array}\right) = \frac{n!}{r!(n-r)!} \tag{3}$$
  counts the number of subsets of size r in a set of n objects and
  $$n! = n \times (n-1) \times (n-2) \times \ldots \times 2 \times 1 \tag{4}$$

2 Conditional probability

- The **conditional probability** of event R given C:
  $$P(R|C) = \frac{P(R \cap C)}{P(C)} \tag{5}$$
  This is the probability of getting an outcome in event R if we know the outcome is in event C.

3 Bayes' theorem

- Two events A and B are said to be **independent** if

$$P(A \cap B) = P(A)P(B) \tag{6}$$

- Two events A and B are **conditionally independent** conditional on a third event C is

$$P(A \cap B|C) = P(A|C)P(B|C) \tag{7}$$

- **Bayes' rule** is

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{8}$$

- In a **naïve Bayes estimator** we estimate $P(X|A, B, \ldots, C)$ by first using Bayes' rule

$$P(X|A, B, \ldots, C) = \frac{P(A, B, \ldots, C|X)P(X)}{P(A, B, \ldots, C)} \tag{9}$$

  and then approximate using an assumption of independence:

$$\begin{aligned} P(A, B, \ldots, C|X) &\approx P(A|X)P(B|X)\ldots P(C|X) \\ P(A, B, \ldots, C) &\approx P(A)P(B)\ldots P(C) \end{aligned} \tag{10}$$

4 Random variables

- A **random variable** is a map from sample space to a set of numerical values.
- The probability that $X = x$, $p(X = x)$, sometimes written $p(x)$, is the sum of the probabilities of all the outcomes with value x.

  1. $$0 \leq p(x) \leq 1 \tag{11}$$
  2. $$\sum_x p(x) = 1 \tag{12}$$

- A **probability distribution** is a table of probabilities for a random variable.
- For two random variables X and Y, the **joint distribution** is $p(x, y)$, the probability $X = x$ and $Y = y$; the **conditional distribution** of $X = x$ given $Y = y$ is $p(x|y)$ and the **marginal distribution** is

$$p(x) = \sum_y p(x, y) \tag{13}$$

- Is $g(x)$ is a function, the **expected value** is

$$\langle g(X) \rangle = \sum_x p(x)g(x) \tag{14}$$

- The **mean** is $\langle X \rangle$. It is often called μ.

- The **variance** is $\langle (X-\mu)^2 \rangle$. It is often called V or σ^2.
- The n**th moment**, often written μ_n, is $\langle X^n \rangle$ and the n**th central moment** is $\langle (X-\mu)^2 \rangle$.
- Expected values have nice properties
  1. $\langle cg(X) \rangle = c\langle g(X) \rangle$
  2. $\langle 1 \rangle = 1$
  3. $\langle g_1(X) + g_2(X) \rangle = \langle g_1(X) \rangle + \langle g_2(X) \rangle$
- Using these nice properties it can be shown that $\sigma^2 = \langle X^2 \rangle - \mu^2$

5 Binomial distribution

- In a **binomial experiment**
  1. There are n identical trials.
  2. Each trial has one of two outcomes, which we call success, S, and failure, F.
  3. The trials are independent.
  4. The random variable of interest, say R, is the total number of successes.
- In a binomial experiment, if p is the chance of success for an individual trial, and $q = 1-p$ is the chance of failure, then the probability of r successes is given by

$$p_R(r) = \begin{pmatrix} n \\ r \end{pmatrix} p^r q^{n-r} \tag{15}$$

- The mean is np and the variance is npq.
- The mean is derived using a fancy trick involving differenciating

$$1 = \sum_{r=0}^{n} \begin{pmatrix} n \\ r \end{pmatrix} p^r q^{n-r} = \sum_{r=0}^{n} p(R=r) \tag{16}$$

  with respect to p.

6 Poisson distribution

- In a **Poisson process** events occur randomly, the rate they occur at doesn't change over time and the chance of an event occuring doesn't depend on when events happened in the past.
- The **Poisson distribution** gives the probability of r events occuring in a time interval if λ is the rate, the average number of events in that period:

$$p(r) = \frac{\lambda^r}{r!} e^{-\lambda} \tag{17}$$

- There is a fancy derivation of this formula which involves subdividing the interval into small subintervals.
- It is possible to show λ is the average count by writing down the formula for the mean and rearranging the terms.

7 The moment generating function

- The **moment generating function** is

$$m(t) = \langle e^{tX} \rangle \tag{18}$$

- It gives the nth moment

$$\frac{d^n m}{dt^n}(0) = \langle X^n \rangle = \mu_n \tag{19}$$

- For the Bernoulli distribution which is one with propability p and zero with probability $q = 1 - p$

$$m(t) = q + pe^t \tag{20}$$

8 Continuous random variables

- The **distribution function** or **cumulative** is

$$F(x) = P(X < x) \tag{21}$$

  so $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to\infty} F(x) = 1$.

- The **density function** is

$$f(x) = \frac{dF}{dx} \tag{22}$$

- By integrating we get

$$F(x) = \int_{-\infty}^{x} f(y)dy \tag{23}$$

  and so

$$\int_{-\infty}^{\infty} f(y)dy = 1 \tag{24}$$

- Hence

$$P(x \in [x_1, x_2]) = F(x_2) - F(x_1) \tag{25}$$

  or

$$P(x \in [x_1, x_2]) = \int_{x_1}^{x_2} f(y)dy \tag{26}$$

- Expected values work much the same way they did for discrete random variables.
- If $Y = X + c$ then $\mu_Y = \mu_x + c$ and $\sigma_Y^2 = \sigma_X^2$.
- If $Y = cX$ then $\mu_Y = c\mu_X$ and $\sigma_Y^2 = c^2\sigma_X^2$.

9 Gauss distribution

- The **Gaußian distribution** has density

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{27}$$

- It has moment generating function

$$m(t) = e^{\mu t + \frac{1}{2}\sigma^2 t^2} \tag{28}$$

  from which can be be shown that the mean is μ and the variance is σ^2 as the notation would suggest.

- To work out probabilities you need to use the **error function**

$$\text{erf}\,(x) = \frac{1}{\sqrt{\pi}} \int_{-x}^{x} e^{-y^2} dy = \frac{2}{\sqrt{\pi}} \int_{0}^{x} e^{-y^2} dy \tag{29}$$

  In fact

$$\text{Prob}(x_1 < x < x_2) = \frac{1}{2}[\text{erf}\,(z_2) - \text{erf}\,(z_1)] \tag{30}$$

  where

$$z = \frac{x - \mu}{\sqrt{2}\sigma} \tag{31}$$

10 Central Limit Theorem

- If X and Y are continuous random variables, with density functions $p_X(x)$ and $p_Y(y)$ and

$$Z = X + Y \tag{32}$$

  then

$$p_Z(z) = \int_{-\infty}^{\infty} p_X(x) p_Y(z - x) dx \tag{33}$$

  This calculation is called a **convolution**.

- If $X \sim \mathcal{N}(\mu_x, \sigma_x^2)$ and $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$ then

$$X + Y = Z \sim \mathcal{N}(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2) \tag{34}$$

- Let $\{X_1, X_2, \ldots, X_n\}$ be a set of random variables. A set of random variables is called **independent identically distributed**, usually abbreviated to i.i.d., if the variables all have the same probability density, say $p_X(x)$ and are independent.

- The **Central limit theorem**: if $\{X_1, X_2, \ldots, X_n\}$ is i.i,d, the **sample mean** is

$$S_n = \frac{X_1 + X_2 + \ldots + X_n}{n} \tag{35}$$

  As n approaches infinity

$$U_n = \sqrt{n}\left(\frac{S_n - \mu}{\sigma}\right) \sim \mathcal{N}(0, 1) \tag{36}$$