Motivation

- Uncertainty is everywhere around us
  - "what is the chance that it will rain today?"
  - "when will the next bus arrive?"
  - "will I go to recitation today?"

- Machine learning tries to understand uncertainties and interact with the real world

- Probability theory is the mathematical study of uncertainty
Basic Concepts

- Sample Space $\Omega$: set of all possible outcomes
- An event, $A$, is a subspace of $\Omega$
- $P(A)$ is the probability of event $A$ occurring
  - $0 \leq P(A) \leq 1$
  - $P(\emptyset) = 0$
  - $P(\Omega) = 1$
  - $P(A^c) = 1 - P(A)$

Examples

Suppose you are rolling a six-sided dice. Then we have:

- $\Omega = \{1, 2, 3, 4, 5, 6\}$
- $P(\{1\}) = P(\{2\}) = P(\{3\}) = P(\{4\}) = P(\{5\}) = P(\{6\}) = \frac{1}{6}$
- $P(\{1, 2, 3, 4, 5, 6\}) = 1$
Given two events, $A$ and $B$, the probability of $A$ or $B$ is

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \quad (1)$$

$$= P(A) + P(B) \quad (2)$$

where (1) and (2) are equal only if $A$ and $B$ are mutually exclusive.

**Examples**

Suppose you are rolling a six-sided dice. Then we have:

- $P(\{2, 4, 6\}) = P(\{2\}) + P(\{4\}) + P(\{6\}) = \frac{1}{2}$

- $P(\{1, 2, 3\} \cup \{2, 4, 6\}) = P(\{1, 2, 3\}) + P(\{2, 4, 6\}) - P(\{2\}) = \frac{1}{2} + \frac{1}{2} - \frac{1}{6} = \frac{5}{6}$
Given two events, $A$ and $B$,

- $\mathbb{P}(A, B)$ is the joint probability of $A$ and $B$ occurring together
- $\mathbb{P}(A \mid B)$ is the conditional probability of $A$ occurring given that we know $B$ has occurred

Joint and Conditional Probabilities are related in the following way:

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A, B)}{\mathbb{P}(B)}$$

where we assume that $\mathbb{P}(B) \neq 0$. 

Examples
Consider a standard deck of cards. What is the probability that the first two cards drawn are both Kings?

- Event $A =$ First card drawn is a King
- Event $B =$ Second card drawn is a King

A standard deck of cards has 4 Kings and 52 total cards.

- $P(A) = \frac{4}{52}$
- $P(B \mid A) = \frac{3}{51}$

This means that

- $P(A, B) = P(B \mid A)P(A) = \frac{3}{51} \cdot \frac{4}{52} = \frac{1}{221} \approx 0.005$
An extension of $\mathbb{P}(A, B) = \mathbb{P}(B \mid A)\mathbb{P}(A)$:

$\mathbb{P}(A_1, A_2, \ldots, A_n) = \mathbb{P}(A_n, \ldots, A_2, A_1)$

$= \mathbb{P}(A_n \mid A_{n-1}, \ldots, A_2, A_1)\mathbb{P}(A_{n-1}, \ldots, A_2, A_1)$

$\vdots$

$= \prod_{i=1}^{n} \mathbb{P}(A_i \mid A_1, A_2, \ldots, A_{i-1})$
Two events, $A$ and $B$, are independent if

$$P(A, B) = P(A)P(B)$$

or equivalently

$$P(A \mid B) = P(A)$$

In other words, knowledge of whether $B$ occurred does not affect the probability that $A$ occurs.
Examples

Suppose that you roll a six-sided die twice. What is the probability that you roll 1 both times?

- $A =$ rolling a 1 in the first roll
- $B =$ rolling a 1 in the second roll

These two events are independent so

$$
\mathbb{P}(A, B) = \mathbb{P}(A)\mathbb{P}(B) = \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36}
$$
Bayes Theorem

We know that the following is true

\[ P(A, B) = P(A \mid B)P(B) = P(B \mid A)P(A) \]

This implies that

\[ P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} \]

\[ P(A \mid B) \propto P(B \mid A)P(A) \]

- \( P(A \mid B) \) - Posterior Probability
- \( P(B \mid A) \) - Likelihood Function
- \( P(A) \) - Prior Information
- \( P(B) \) - Evidence
If a person has an allergy (A), sneezing (S) is observed with probability $P(S \mid A) = 0.8$. What is the chance a person has an allergy given that they are sneezing: $P(A \mid S)$?

- Assume that $P(A) = 0.001$ (few people have allergies)
- Assume that $P(S) = 0.1$ (many people sneeze).

\[
P(A \mid S) = \frac{P(S \mid A)P(A)}{P(S)} = \frac{0.8 \cdot 0.001}{0.1} = 0.008
\]
Random Variables

So far, we have only considered simple examples with simple events as the possible outcomes

- $A =$ rolling 1 on a six-sided dice
- $B =$ first card drawn is a King

This is mathematically imprecise and can be limiting. The notion of random variables resolves these issues

- A random variable $X$ is a function $X : \Omega \rightarrow \mathbb{R}$
- Abuse of notation: $P(X = x)$ is the probability that the random variable takes on value $x$
Random Variables

**Examples**

- Dice Roll: $X = i$ for $i \in \{1, 2, 3, 4, 5, 6\}$ with $\mathbb{P}(X = i) = \frac{1}{6}$
- Biased Coin: $X = 1$ with $\mathbb{P}(X = 1) = p$ and $X = 0$ with $\mathbb{P}(X = 0) = 1 - p$

All random variables have a Cumulative Distribution Function (CDF):

$$F(x) = \mathbb{P}(X \leq x)$$

Some properties of the CDF are:

- $0 \leq F(x) \leq 1$
- $F(X)$ is monotonically increasing
- $\lim_{x \to -\infty} F(x) = 0$  \hspace{1cm} $\lim_{x \to \infty} F(x) = 1$
Discrete Random Variables

A random variable that can take on only finitely many different values

Discrete random variables have a Probability Mass Function (PMF):

\[ p(x) = P(X = x) \]

The PMF satisfies

\[ \sum_x P(X = x) = 1 \]

Examples

Suppose you are flipping two coins. Let \( X \) be the random variable that equals the number of heads

\[
\begin{array}{|c|c|c|}
\hline
X & P(X = 0) & P(X = 1) & P(X = 2) \\
\hline
0  & 0.25   & 0.5    & 0.25   \\
\hline
\end{array}
\]
Continuous Random Variables

A random variable that can take on infinitely many different values.

Continuous random variables have a Probability Distribution Function (PDF):

\[ p(x) = f(x) = \frac{d}{dx} F(x) \]

The PDF satisfies

\[ \int_{-\infty}^{\infty} f(x) \, dx = 1 \]

Examples

Suppose that \( X \) is a random variable that is equal to a value chosen uniformly at random from the interval \([0, a]\). Then

\[ F(x) = \frac{x}{a} \quad f(x) = \frac{1}{a} \]
Suppose that for two random variables, $X$ and $Y$, the joint distribution $p(x, y)$ is known for all combinations of $X$ and $Y$. Then the marginal distribution of $X$ is

$$p(x) = \sum_y p(x, y)$$

and the marginal distribution of $Y$ is

$$p(y) = \sum_x p(x, y)$$

where

$$p(x) = \int_{-\infty}^{\infty} p(x, y) dy$$

and

$$p(y) = \int_{-\infty}^{\infty} p(x, y) dx$$
Consider two random variables $X$ and $Y$. Suppose the joint distribution is given by

$$
\begin{array}{c|cccc|c}
 & x_1 & x_2 & x_3 & x_4 & p(y) \\
\hline
y_1 & \frac{4}{32} & \frac{2}{32} & \frac{1}{32} & \frac{1}{32} & \frac{8}{32} \\
y_2 & \frac{2}{32} & \frac{4}{32} & \frac{1}{32} & \frac{1}{32} & \frac{8}{32} \\
y_3 & \frac{2}{32} & \frac{2}{32} & \frac{2}{32} & \frac{8}{32} \\
y_4 & \frac{8}{32} & 0 & 0 & \frac{8}{32} \\
p(x) & \frac{16}{32} & \frac{4}{32} & \frac{4}{32} & \frac{32}{32} \\
\end{array}
$$

$$
p(x_1) = \sum_y p(x_1, y) = \frac{16}{32}
$$
Expected Value

The expected value of a random variable, \( X \), is the mean of the distribution

\[
E[X] = \sum_x x \mathbb{P}(X = x) \quad \quad \quad \quad \quad E[X] = \int_x x f(x) \, dx
\]

Expectation is linear

- \( E[aX + b] = aE[X] + b \)
- \( E[X + Y] = E[X] + E[Y] \)

Examples

Suppose that \( X \) and \( Y \) are random variables that are 1 with probability \( p \) and 0 otherwise. What is the expected value of \( X + Y \)?

\[
E[X + Y] = E[X] + E[Y] = \mathbb{P}(X = 1) + \mathbb{P}(Y = 1) = 2p
\]
The variance of a random variable is the measure of the 'spread' of the values the variable takes on:

$$Var(X) = \mathbb{E}[(X - \mathbb{E}[X])^2]$$

The variance can also be written as

$$Var(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

Variance is generally not linear

- $Var(aX + b) = a^2 \ Var(X)$
- $Var(X + Y) = Var(X) + Var(Y) + 2 \text{Cov}(X, Y)$
Covariance

The covariance between random variables $X$ and $Y$ measures the degree to which $X$ and $Y$ are related

$$\text{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$$

This can also be written as

$$\text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

Note that $\text{Cov}(X, X) = \text{Var}(X)$
Examples

Suppose that $X$ and $Y$ are random variables that are 1 with probability $p$ and 0 otherwise. What is the covariance of $X$ and $Y$?

What is the variance of $X + Y$?

$$\text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

$$= p^2(1) + (1 - p^2)(0) - p^2 = 0$$

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y)$$

$$= \mathbb{E}[X^2] - \mathbb{E}[X]^2 + \mathbb{E}[Y^2] - \mathbb{E}[Y]^2$$

$$= p - p^2 + p - p^2 = 2p(1 - p)$$
Important Discrete Distributions

- **Bernoulli \((p)\)**
  - \(\mathbb{P}(X = x) = p^x(1 - p)^{1-x}\) for \(x \in \{0, 1\}\)
  - \(\mathbb{E}[x] = p\)

- **Binomial \((n, p)\)**
  - \(\mathbb{P}(X = x) = \binom{n}{x} p^x(1 - p)^{n-x}\)
  - \(\mathbb{E}[x] = np\)

- **Geometric \((p)\)**
  - \(\mathbb{P}(X = x) = p(1 - p)^{x-1}\)
  - \(\mathbb{E}[x] = \frac{1}{p}\)

- **Poisson \((\lambda)\)**
  - \(\mathbb{P}(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}\)
  - \(\mathbb{E}[x] = \lambda\)
Important Discrete Distributions

Examples

- Bernoulli - flipping a biased coin that lands heads with probability $p$
- Binomial - flipping $n$ biased coins that land heads with probability $p$
- Geometric - flipping a biased coin that lands heads with probability $p$ until it lands on heads
- Poisson - letters received in a given day when the average letters received per day is $\lambda$
Important Continuous Distributions

- **Uniform** \((a, b)\) with \(a \leq b\)
  - \(f(x) = \frac{1}{b-a}\) for \(a \leq x \leq b\), 0 otherwise
  - \(\mathbb{E}[x] = \frac{a+b}{2}\)

- **Normal** \((\mu, \sigma^2)\)
  - \(f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}\)
  - \(\mathbb{E}(x) = \mu\)

**Examples**
- Uniform - spinning a board game spinner
- Normal - height of a person in the general population
Let $X_1, X_2, \ldots, X_n$ be a set of independent and identically distributed (iid) random variables with $\mathbb{E}[X_i] = \mu$. Then the sample average given by

$$\bar{X}_n = \frac{1}{n}(X_1 + X_1 + \ldots + X_n)$$

converges to the expected value

$$\bar{X}_n \rightarrow \mu \text{ as } n \rightarrow \infty$$
Let $X_1, X_2, \ldots, X_n$ be a set of independent and identically distributed (iid) random variables with $\mathbb{E}[X_i] = \mu$ and $\text{Var}(X_i) = \sigma^2$, both finite. If we define the random variable $Y$ as

$$Y = \frac{1}{n}(X_1 + X_1 + \ldots + X_n)$$

then $Y$ is approximately normal with mean $\mu$ and variance $\frac{\sigma^2}{n}$ \//

The approximation improves as $n \to \infty$
Sampling is the process by which elements are drawn from a population or distribution.

In this course, we typically assume elements are sampled independently (i.e. with replacement).

The `np.random` module contains many functions for sampling randomly from various distributions.

```python
np.random.uniform(0,1)  # uniform on [0,1]
```
The importance of the iid assumption cannot be overstated!

Independence assumption simplifies many things:

\[ P(X, Y) = P(X)P(Y) \]
\[ P(X \mid Y) = P(X) \]
\[ Var(X + Y) = Var(X) + Var(Y) \]
Parameter Estimation

Suppose we have a parametrized distribution $\mathbb{P}(X; \theta)$ with $\theta$ unknown.

Suppose we have an iid set of samples $x_1, \ldots, x_n$.

How can we estimate $\theta$? From Bayes’ Theorem, we have that

$$
\mathbb{P}(\theta \mid X) \propto \mathbb{P}(X \mid \theta)\mathbb{P}(\theta)
$$

- **MAP:** $\hat{\theta} = \arg \max_{\theta} \mathbb{P}(\theta \mid X)$
- **MLE:** $\hat{\theta} = \arg \max_{\theta} \mathbb{P}(X \mid \theta)$
The logarithmic function is monotonically increasing - will not change the location of where a function achieves its maximum

Multiplication turns into summation, resulting in less overflow

Considering the negative of the likelihood function allows us to use gradient descent to minimize the function

\[
\arg \max_{\theta} f(X \mid \theta) = \arg \min_{\theta} - \log f(X \mid \theta)
\]
Parameter Estimation - Binomial Distribution

Examples

Suppose you toss a (possibly biased) coin $n$ times and record the number of heads, $k$. What is $p$, the probability that it lands heads on a given coin flip. We can use the binomial distribution and MLE to find $p$:

$$
\hat{p} = \arg \max_p \mathbb{P}(k \mid p) = \arg \max_p \binom{n}{k} p^k (1-p)^{n-k}
$$

$$
= \arg \max_p p^k (1-p)^{n-k}
$$

$$
= \arg \min_p -\log p^k (1-p)^{n-k}
$$

$$
= \arg \min_p -k \log p - (n-k) \log (1-p)
$$

Taking the derivative wrt to $p$ and zeroing implies that

$$
\hat{p} = \frac{k}{n}
$$
MLE with Logistic Regression

**Examples**

Suppose you have an iid dataset, \( S = \{(x_i, y_i)_{i=1}^N\} \), and want to model it using a "log-linear" model:

\[
\mathbb{P}(y | x, w, b) = \frac{1}{1 + e^{-y(w^T x + b)}}
\]

We can use a MLE to find \( w, b \):

\[
\hat{w}, \hat{b} = \arg \max_{w, b} \mathbb{P}(y | x, w, b)
\]

\[
= \arg \max_{w, b} \prod_{i=1}^{N} \mathbb{P}(y_i | x_i, w, b) \quad \text{iid assumption}
\]

\[
= \arg \min_{w, b} \sum_{i=1}^{N} - \ln \mathbb{P}(y_i | x_i, w, b) \quad \text{log-loss in logistic reg.}
\]
References

- Lucy Yin’s Recitation on Probability (CS155 Winter 2016)
- Kevin Murphy’s Machine Learning: A Probabilistic Perspective