Probability and Sampling Recitation

Avishek Dutta

CS155 Machine Learning and Data Mining

February 2, 2017

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 Ω: set of all possible outcomes
- An event, A, is a subspace of Ω
- $\mathbb{P}(A)$ is the probability of event A occurring
  - $0 \leq \mathbb{P}(A) \leq 1$
  - $\mathbb{P}(\emptyset) = 0$
  - $\mathbb{P}(\Omega) = 1$
  - $\mathbb{P}(\overline{A}) = 1 - \mathbb{P}(A)$

Examples

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

- $\Omega = \{1, 2, 3, 4, 5, 6\}$
- $\mathbb{P}(\{1\}) = \mathbb{P}(\{2\}) = \mathbb{P}(\{3\}) = \mathbb{P}(\{4\}) = \mathbb{P}(\{5\}) = \mathbb{P}(\{6\}) = \frac{1}{6}$
- $\mathbb{P}(\{1, 2, 3, 4, 5, 6\}) = 1$

Union

Given two events, A and B, the probability of A or B is

$$\begin{aligned} \mathbb{P}(A \cup B) &= \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B) && (1) \\ &= \mathbb{P}(A) + \mathbb{P}(B) && (2) \end{aligned}$$

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:

- $\mathbb{P}(\{2,4,6\}) = \mathbb{P}(\{2\}) + \mathbb{P}(\{4\}) + \mathbb{P}(\{6\}) = \frac{1}{2}$
- $\mathbb{P}(\{1,2,3\} \cup \{2,4,6\}) =$
  $\mathbb{P}(\{1,2,3\}) + \mathbb{P}(\{2,4,6\}) - \mathbb{P}(\{2\}) = \frac{1}{2} + \frac{1}{2} - \frac{1}{6} = \frac{5}{6}$

Joint and Conditional Probabilities

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 $\mathbb{P}(B) \neq 0$.

Joint and Conditional Probabilities

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.

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

This means that

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

Joint Probabilities - Chain Rule

An extension of $\mathbb{P}(A, B) = \mathbb{P}(B \mid A)\mathbb{P}(A)$:

$$\begin{aligned}
\mathbb{P}(A_1, A_2, \dots, A_n) &= \mathbb{P}(A_n, \dots, A_2, A_1) \\
&= \mathbb{P}(A_n \mid A_{n-1}, \dots, A_2, A_1)\mathbb{P}(A_{n-1}, \dots, A_2, A_1) \\
&\vdots \\
&= \prod_{i=1}^{n} \mathbb{P}(A_i \mid A_1, A_2, \dots, A_{i-1})
\end{aligned}$$

Independence

Two events, A and B, are independent if

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

or equivalently

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

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

Independence

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

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

This implies that

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

$$\mathbb{P}(A \mid B) \propto \mathbb{P}(B \mid A)\mathbb{P}(A)$$

- $\mathbb{P}(A \mid B)$ - Posterior Probability
- $\mathbb{P}(B \mid A)$ - Likelihood Function
- $\mathbb{P}(A)$ - Prior Information
- $\mathbb{P}(B)$ - Evidence

Bayes Theorem

Examples

If a person has an allergy (A), sneezing (S) is observed with probability $\mathbb{P}(S \mid A) = 0.8$. What is the chance a person has an allergy given that they are sneezing: $\mathbb{P}(A \mid S)$?

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

$$\begin{aligned}\mathbb{P}(A \mid S) &= \frac{\mathbb{P}(S \mid A)\mathbb{P}(A)}{\mathbb{P}(S)} \\ &= \frac{0.8 \cdot 0.001}{0.1} \\ &= 0.008\end{aligned}$$

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 $X : \Omega \to \mathbb{R}$
- Abuse of notation: $\mathbb{P}(X = x)$ 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

$$\lim_{x \to -\infty} F(x) = 0 \qquad \lim_{x \to \infty} F(x) = 1$$

Discrete Random Variables

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

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

$$p(x) = \mathbb{P}(X = x)$$

The PMF satisfies

$$\sum_x \mathbb{P}(X = x) = 1$$

Examples

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

$\mathbb{P}(X = 0)$	$\mathbb{P}(X = 1)$	$\mathbb{P}(X = 2)$
0.25	0.5	0.25

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

$$F(x) = \frac{x}{a} \qquad f(x) = \frac{1}{a}$$

Marginal Distribution

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

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

and the marginal distribution of Y is

$$p(y) = \sum_x p(x, y) \qquad p(y) = \int_{-\infty}^{\infty} p(x, y)dx$$

Marginal Distribution

Examples

Consider two random variables X and Y. Suppose the joint distribution is given by

	x_1	x_2	x_3	x_4	$p(y)$
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	$\frac{2}{32}$	$\frac{2}{32}$	$\frac{2}{32}$	$\frac{2}{32}$	$\frac{8}{32}$
y_4	$\frac{8}{32}$	0	0	0	$\frac{8}{32}$
$p(x)$	$\frac{16}{32}$	$\frac{8}{32}$	$\frac{4}{32}$	$\frac{4}{32}$	$\frac{32}{32}$

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

Expected Value

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

$$\mathbb{E}[X] = \sum_x x\mathbb{P}(X = x) \qquad\qquad \mathbb{E}[X] = \int_x xf(x)dx$$

Expectation is linear

- $\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$
- $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$

Examples

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

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y] = \mathbb{P}(X = 1) + \mathbb{P}(Y = 1) = 2p$$

Variance

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

- $Var(aX + b) = a^2 Var(X)$
- $Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y)$

Covariance

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

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

This can also be written as

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

Note that $Cov(X, X) = Var(X)$

Variance and Covariance

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 $X + Y$?

$$\begin{aligned}
Cov(X, Y) &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \\
&= p^2(1) + (1 - p^2)(0) - p^2 = 0 \\
Var(X + Y) &= Var(X) + Var(Y) + 2Cov(X, Y) \\
&= \mathbb{E}[X^2] - \mathbb{E}[X]^2 + \mathbb{E}[Y^2] - \mathbb{E}[Y]^2 \\
&= p - p^2 + p - p^2 = 2p(1 - p)
\end{aligned}$$

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 (λ)
  - $\mathbb{P}(X = x) = \frac{\lambda^x e^{-x}}{x!}$
  - $\mathbb{E}[x] = \lambda$

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 λ

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 (μ, σ^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

Law of Large Numbers

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

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

converges to the expected value

$$\overline{X}_n \to \mu \text{ as } n \to \infty$$

Central Limit Theorem

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 $Var(X_i) = \sigma^2$, 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 μ and variance $\frac{\sigma^2}{n}$ //

The approximation improves as $n \to \infty$

Sampling

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

```
np.random.uniform(0,1)          # uniform on [0,1]
```

Independent and Identically Distributed

The importance of the iid assumption cannot be overstated!

Independence assumption simplifies many things:

$$\mathbb{P}(X, Y) = \mathbb{P}(X)\mathbb{P}(Y)$$

$$\mathbb{P}(X \mid Y) = \mathbb{P}(X)$$

$$Var(X + Y) = Var(X) + Var(Y)$$

Parameter Estimation

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

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

How can we estimate θ? 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: $\hat{\theta} = \arg\max_{\theta} \mathbb{P}(X \mid \theta)$

Parameter Estimate - log trick

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:

$$
\begin{aligned}
\hat{p} = \arg\max_p \mathbb{P}(k \mid p) &= \arg\max_p \binom{n}{k} p^k (1-p)^{n-k} \\
&= \arg\max_p p^k (1-p)^{n-k} \\
&= \arg\min_p -\log p^k (1-p)^{n-k} \\
&= \arg\min_p -k \log p - (n-k) \log (1-p)
\end{aligned}
$$

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:

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

We can use a MLE to find w, b:

$$\begin{aligned}
\hat{w}, \hat{b} &= \arg\max_{w,b} \mathbb{P}(y \mid x, w, b) \\
&= \arg\max_{w,b} \prod_{i=1}^{N} \mathbb{P}(y_i \mid x_i, w, b) && \text{iid assumption} \\
&= \arg\min_{w,b} \sum_{i=1}^{N} -\ln \mathbb{P}(y_i \mid x_i, w, b) && \text{log-loss in logistic reg.}
\end{aligned}$$

References

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