Probability Theory for Machine Learning

Jesse Bettencourt
September 2017
Introduction to Machine Learning
CSC411
University of Toronto
Introduction to Notation
Motivation

Uncertainty arises through:

- Noisy measurements
- Finite size of data sets
- Ambiguity
- Limited Model Complexity

Probability theory provides a consistent framework for the quantification and manipulation of uncertainty.
Sample space $\Omega$ is the set of all possible outcomes of an experiment.

Observations $\omega \in \Omega$ are points in the space also called sample outcomes, realizations, or elements.

Events $E \subset \Omega$ are subsets of the sample space.
In this experiment we flip a coin twice:

**Sample space** All outcomes $\Omega = \{HH, HT, TH, TT\}$

**Observation** $\omega = HT$ valid sample since $\omega \in \Omega$

**Event** Both flips same $E = \{HH, TT\}$ valid event since $E \subset \Omega$
Probability
The probability of an event $E$, $P(E)$, satisfies three axioms:

1: $P(E) \geq 0$ for every $E$

2: $P(\Omega) = 1$

3: If $E_1, E_2, \ldots$ are disjoint then

\[ P\left( \bigcup_{i=1}^{\infty} E_i \right) = \sum_{i=1}^{\infty} P(E_i) \]
Joint Probability of $A$ and $B$ is denoted $P(A, B)$

Conditional Probability of $A$ given $B$ is denoted $P(A|B)$.

- Assuming $P(B) > 0$, then $P(A|B) = P(A, B)/P(B)$
- Product Rule: $P(A, B) = P(A|B)P(B) = P(B|A)P(A)$
60% of ML students pass the final and 45% of ML students pass both the final and the midterm. What percent of students who passed the final also passed the midterm?
60% of ML students pass the final and 45% of ML students pass both the final and the midterm.
What percent of students who passed the final also passed the midterm?
Reword: What percent passed the midterm given they passed the final?

\[ P(M|F) = P(M, F)/P(F) \]
\[ = 0.45/0.60 \]
\[ = 0.75 \]
Events $A$ and $B$ are independent if $P(A, B) = P(A)P(B)$

Events $A$ and $B$ are conditionally independent given $C$ if $P(A, B|C) = P(B|A, C)P(A|C) = P(B|C)P(A|C)$
Marginalization and Law of Total Probability

Marginalization (Sum Rule)

\[ P(X) = \sum_{Y} P(X, Y) \]

Law of Total Probability

\[ P(X) = \sum_{Y} P(X|Y)P(Y) \]
Bayes’ Rule
Bayes’ Rule:

\[ P(A|B) = \frac{P(B|A)P(A)}{P(B)} \]
Bayes’ Rule:

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

\[ P(\theta|x) = \frac{P(x|\theta)P(\theta)}{P(x)} \]

Posterior = Likelihood \times Prior

\[ Posterior \propto Likelihood \times Prior \]
Suppose you have tested positive for a disease. What is the probability you actually have the disease?
Bayes’ Example

Suppose you have tested positive for a disease. What is the probability you actually have the disease?
This depends on accuracy and sensitivity of test and prior probability of the disease:

\[ P(T = 1|D = 1) = 0.95 \text{ (true positive)} \]
\[ P(T = 1|D = 0) = 0.10 \text{ (false positive)} \]
\[ P(D = 1) = 0.1 \text{ (prior)} \]

So \[ P(D = 1|T = 1) = ? \]
Suppose you have tested positive for a disease. What is the probability you actually have the disease?

\[ P(T = 1|D = 1) = 0.95 \text{ (true positive)} \]

\[ P(T = 1|D = 0) = 0.10 \text{ (false positive)} \]

\[ P(D = 1) = 0.1 \text{ (prior)} \]

So \( P(D = 1|T = 1) = ? \)

Use Bayes’ Rule:

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

\[
P(D = 1|T = 1) = \frac{P(T = 1|D = 1)P(D = 1)}{P(T = 1)}
\]
Bayes’ Example

Suppose you have tested positive for a disease. What is the probability you actually have the disease?

\[ P(T = 1|D = 1) = 0.95 \text{ (true positive)} \]
\[ P(T = 1|D = 0) = 0.10 \text{ (false positive)} \]
\[ P(D = 1) = 0.1 \text{ (prior)} \]

Use Bayes’ Rule:

\[ P(D = 1|T = 1) = \frac{P(T = 1|D = 1)P(D = 1)}{P(T = 1)} \]
\[ P(D = 1|T = 1) = \frac{0.95 \times 0.1}{P(T = 1)} \]
Suppose you have tested positive for a disease. What is the probability you actually have the disease?

\[
P(T = 1|D = 1) = 0.95 \text{ (true positive)}
\]
\[
P(T = 1|D = 0) = 0.10 \text{ (false positive)}
\]
\[
P(D = 1) = 0.1 \text{ (prior)}
\]

By Bayes' Rule

\[
P(D = 1|T = 1) = \frac{0.95 \times 0.1}{P(T = 1)}
\]

By Law of Total Probability

\[
P(T = 1) = \sum_D P(T = 1|D)P(D)
\]
\[
= P(T = 1|D = 1)P(D = 1) + P(T = 1|D = 0)P(D = 0)
\]
\[
= 0.95 \times 0.1 + 0.1 \times 0.90
\]
\[
= 0.185
\]
Suppose you have tested positive for a disease. What is the probability you actually have the disease?

\[
P(T = 1|D = 1) = 0.95 \text{ (true positive)}
\]

\[
P(T = 1|D = 0) = 0.10 \text{ (false positive)}
\]

\[
P(D = 1) = 0.1 \text{ (prior)}
\]

\[
P(T = 1) = 0.185 \text{ (from Law of Total Probability)}
\]

\[
P(D = 1|T = 1) = \frac{0.95 \times 0.1}{P(T = 1)} = \frac{0.95 \times 0.1}{0.185} = 0.51
\]

Probability you have the disease given you tested positive is 51%
Random Variables and Statistics
How do we connect sample spaces and events to data?

A random variable is a mapping which assigns a real number $X(\omega)$ to each observed outcome $\omega \in \Omega$.

For example, let's flip a coin 10 times. $X(\omega)$ counts the number of Heads we observe in our sequence. If $\omega = HHTHTHHTHT$ then $X(\omega) = 6$. 
Random variables are said to be independent and identically distributed (i.i.d.) if they are sampled from the same probability distribution and are mutually independent. This is a common assumption for observations. For example, coin flips are assumed to be iid.
Discrete and Continuous Random Variables

Discrete Random Variables

- Takes countably many values, e.g., number of heads
- Distribution defined by probability mass function (PMF)
- Marginalization: $p(x) = \sum_y p(x, y)$

Continuous Random Variables

- Takes uncountably many values, e.g., time to complete task
- Distribution defined by probability density function (PDF)
- Marginalization: $p(x) = \int_y p(x, y)dy$
**Mean:** First Moment, $\mu$

\[
E[x] = \sum_{i=1}^{\infty} x_i p(x_i) \quad \text{(univariate discrete r.v.)}
\]

\[
E[x] = \int_{-\infty}^{\infty} xp(x) \, dx \quad \text{(univariate continuous r.v.)}
\]

**Variance:** Second Moment, $\sigma^2$

\[
Var[x] = \int_{-\infty}^{\infty} (x - \mu)^2 p(x) \, dx
\]

\[
= E[(x - \mu)^2]
\]

\[
= E[x^2] - E[x]^2
\]
Gaussian Distribution
Univariate Gaussian Distribution

Also known as the Normal Distribution, \( N(\mu, \sigma^2) \)

\[
N(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi \sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x - \mu)^2\right\}
\]
Multivariate Gaussian Distribution

Multidimensional generalization of the Gaussian.

\( \mathbf{x} \) is a \( D \)-dimensional vector

\( \mu \) is a \( D \)-dimensional mean vector

\( \Sigma \) is a \( D \times D \) covariance matrix with determinant \( |\Sigma| \)

\[
\mathcal{N}(\mathbf{x}|\mu, \Sigma) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} \exp\left\{ -\frac{1}{2} (\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu) \right\}
\]
Recall that $\mathbf{x}$ and $\mu$ are $D$-dimensional vectors.

**Covariance matrix** $\Sigma$ is a matrix whose $(i,j)$ entry is the covariance

$$
\Sigma_{ij} = \text{Cov}(\mathbf{x}_i, \mathbf{x}_j) \\
= E[(\mathbf{x}_i - \mu_i)(\mathbf{x}_j - \mu_j)] \\
= E[(\mathbf{x}_i\mathbf{x}_j)] - \mu_i\mu_j
$$

so notice that the diagonal entries are the variance of each element.

The covariant matrix has the property that it is symmetric and positive-semidefinite (this is useful for whitening).
Whitening is a linear transform that converts a $d$-dimensional random vector $\mathbf{x} = (x_1, \ldots, x_d)^T$ with mean $\mu = E[\mathbf{x}] = (\mu_1, \ldots, \mu_d)^T$ and positive definite $d \times d$ covariance matrix $\text{Cov}(\mathbf{x}) = \Sigma$ into a new random $d$-dimensional vector $\mathbf{z} = (z_1, \ldots, z_d)^T = \mathbf{Wx}$ with “white” covariance matrix, $\text{Cov}(\mathbf{z}) = \mathbf{I}$.

The $d \times d$ covariance matrix $\mathbf{W}$ is called the whitening matrix. Mahalanobis or ZCA whitening matrix: $\mathbf{W}_{\text{ZCA}} = \Sigma^{-\frac{1}{2}}$
Inferring Parameters
Inferring Parameters

We have data \( X \) and we assume it is sampled from some distribution. How do we figure out the parameters that ‘best’ fit that distribution?

Maximum Likelihood Estimation (MLE)

\[
\hat{\theta}_{MLE} = \arg\max_{\theta} P(X|\theta)
\]

Maximum a Posteriori (MAP)

\[
\hat{\theta}_{MAP} = \arg\max_{\theta} P(\theta|X)
\]
MLE for Univariate Gaussian Distribution

We are trying to infer the parameters for a Univariate Gaussian Distribution, mean ($\mu$) and variance ($\sigma^2$).

$$
\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x - \mu)^2\right\}
$$

The **likelihood** that our observations $x_1, \ldots, x_N$ were generated by a univariate Gaussian with parameters $\mu$ and $\sigma^2$ is

$$
\text{Likelihood} = p(x_1 \ldots x_N|\mu, \sigma^2) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x_i - \mu)^2\right\}
$$
MLE for Univariate Gaussian Distribution

For MLE we want to maximize this likelihood, which is difficult because it is represented by a product of terms

$$\text{Likelihood} = p(x_1 \ldots x_N | \mu, \sigma^2) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2} (x_i - \mu)^2 \right\}$$

So we take the log of the likelihood so the product becomes a sum

$$\log \text{Likelihood} = \log p(x_1 \ldots x_N | \mu, \sigma^2)$$

$$= \sum_{i=1}^{N} \log \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2} (x_i - \mu)^2 \right\}$$

Since log is monotonically increasing $\max L(\theta) = \max \log L(\theta)$
MLE for Univariate Gaussian Distribution

The log Likelihood simplifies to

\[ \mathcal{L}(\mu, \sigma) = \sum_{i=1}^{N} \log \left( \frac{1}{\sqrt{2\pi}\sigma^2} \right) \exp\left\{ -\frac{1}{2\sigma^2} (x_i - \mu)^2 \right\} \]

\[ = -\frac{1}{2} N \log(2\pi\sigma^2) - \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{2\sigma^2} \]

Which we want to maximize. How?
MLE for Univariate Gaussian Distribution

To maximize we take the derivatives, set equal to 0, and solve:

\[
\mathcal{L}(\mu, \sigma) = -\frac{1}{2} N \log(2\pi\sigma^2) - \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{2\sigma^2}
\]

Derivative w.r.t. \( \mu \), set equal to 0, and solve for \( \hat{\mu} \)

\[
\frac{\partial \mathcal{L}(\mu, \sigma)}{\partial \mu} = 0 \implies \hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} x_i
\]

Therefore the \( \hat{\mu} \) that maximizes the likelihood is the average of the data points.

Derivative w.r.t. \( \sigma^2 \), set equal to 0, and solve for \( \hat{\sigma}^2 \)

\[
\frac{\partial \mathcal{L}(\mu, \sigma)}{\partial \sigma^2} = 0 \implies \hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \hat{\mu})^2
\]