Introduction to Machine Learning

Probability
I Foundations of Murphy book

Mingchen Gao

Computer Science & Engineering
State University of New York at Buffalo
Buffalo, NY, USA
mgao8@buffalo.edu
Slides Adapted from Varun Chandola

University at Buffalo
Department of Computer Science and Engineering
School of Engeering and Applied Sciences

Outline

What is Probability? [3, 1]

- Probability that a coin will land heads is 50%[1]
- **What does this mean?**

[1] Dr. Persi Diaconis showed that a coin is 51% likely to land facing the same way up as it is started.


FREQUENTISTS
BAYESIANS

Frequentist Interpretation

- Number of times an event will be observed in *n trials*
- What if the event can only occur once?
  - *My winning the next month's powerball.*
  - *Polar ice caps melting by year 2050.*

Bayesian Interpretation

- **Uncertainty** of the event
- Use for making decisions
  - What is the probability of an email is spam?

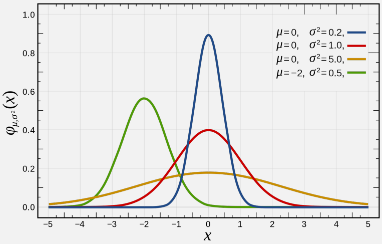
What is a Random Variable (X)?

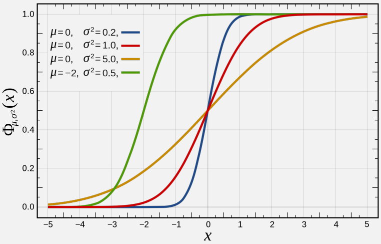
- X - random variable ($\mathbf{X}$ if multivariate)
- x - a specific value taken by the random variable (($\mathbf{x}$ if multivariate))
- Can take any value from $\mathcal{X}$
- **Discrete Random Variable** - $\mathcal{X}$ is finite/countably finite
- **Continuous Random Variable** - $\mathcal{X}$ is infinite
- $P(X = x)$ or $P(x)$ is the probability of X taking value x
- $p(x)$ is either the **probability mass function** (discrete) or **probability density function** (continuous) for the random variable X at x

Discrete Examples

1. Coin toss ($\mathcal{X} = \{heads, tails\}$)
2. Six sided dice ($\mathcal{X} = \{1, 2, 3, 4, 5, 6\}$)

Continuous Example: Gaussian distribution





Basic Rules - Quick Review

- For two events A and B:
  - **Union of two events**
    - $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$
  - **Joint Probability**
    - **product rule** $P(A, B) = P(A \wedge B) = P(A|B)P(B)$
    - **sum rule**
      Given $P(A, B)$ what is $P(A)$?
      Sum $P(A, B)$ over all values for B

$$P(A) = \sum_b P(A, B) = \sum_b P(A|B = b)P(B = b)$$

Basic Rules - Quick Review

- **Chain Rule of Probability**
  - Given D random variables, $\{X_1, X_2, \ldots, X_D\}$

$$P(X_{1:D}) = P(X_1)P(X_2|X_1)P(X_3|X_1, X_2)\ldots P(X_D|X_{1:D-1})$$

- **Conditional Probability**
  - $P(A|B) = \frac{P(A,B)}{P(B)}$

Bayes Rule or Bayes Theorem

- Computing $P(X = x|Y = y)$:

Bayes Theorem

$$\begin{aligned}
P(X = x|Y = y) &= \frac{P(X = x, Y = y)}{P(Y = y)} \\
&= \frac{P(X = x)P(Y = y|X = x)}{\sum_{x'} P(X = x')P(Y = y|X = x')}
\end{aligned}$$

Example

- Medical Diagnosis
- Random event 1: A *test* is positive or negative (X)
- Random event 2: A person has cancer (Y) – yes or no
- What we know:
  1. Test has accuracy of 80%
  2. Number of times the test is positive when the person has cancer

$$P(X = 1|Y = 1) = 0.8$$

  3. Prior probability of having cancer is 0.4%

$$P(Y = 1) = 0.004$$

Question?

If I test positive, does it mean that I have 80% rate of cancer?

Base Rate Fallacy

- Ignored the prior information
- What we need is:

$$P(Y = 1|X = 1) =?$$

- More information:
  - False positive (alarm) rate for the test
  - $P(X = 1|Y = 0) = 0.1$

$$P(Y = 1|X = 1) = \frac{P(X = 1|Y = 1)P(Y = 1)}{P(X = 1|Y = 1)P(Y = 1) + P(X = 1|Y = 0)P(Y = 0)}$$

Classification Using Bayes Rules

- Given input example $\mathbf{X}$, Y is the random variable denoting the true class, we want to find the true class

$$P(Y = c|\mathbf{X})$$

- Assuming the **class-conditional** probability $P(\mathbf{X}|Y = c)$ and class prior $P(Y = c)$ are known
- Applying Bayes Rule

$$P(Y = c|\mathbf{X}) = \frac{P(Y = c)P(\mathbf{X}|Y = c)}{\sum_c P(Y = c'))P(\mathbf{X}|Y = c')}$$

Independence

- One random variable does not depend on another
- $A \perp B \iff P(A, B) = P(A)P(B)$
- Joint written as a product of marginals
- **Conditional Independence**

$$A \perp B|C \iff P(A, B|C) = P(A|C)P(B|C)$$

A is conditionally independent of B given C

More About Conditional Independence

- Alice and Bob live in the same town but far away from each other
- Alice drives to work and Bob takes the bus
- Event A - Alice comes late to work
- Event B - Bob comes late to work
- Event C - A snow storm has hit the town
- $P(A|C)$ - Probability that Alice comes late to work given there is a snowstorm

More About Conditional Independence

- Alice and Bob live in the same town but far away from each other
- Alice drives to work and Bob takes the bus
- Event A - Alice comes late to work
- Event B - Bob comes late to work
- Event C - A snow storm has hit the town
- $P(A|C)$ - Probability that Alice comes late to work given there is a snowstorm
- Now if I also know that Bob has come late to work, will it change the probability that Alice comes late to work?

More About Conditional Independence

- Alice and Bob live in the same town but far away from each other
- Alice drives to work and Bob takes the bus
- Event A - Alice comes late to work
- Event B - Bob comes late to work
- Event C - A snow storm has hit the town
- $P(A|C)$ - Probability that Alice comes late to work given there is a snowstorm
- Now if I also know that Bob has come late to work, will it change the probability that Alice comes late to work?
- What if I do not observe C? Will B have any impact on probability of A happening?

Continuous Random Variables

- X is continuous
- Can take any value
- How does one define probability?
  - $\sum_x P(X = x) = 1$

Continuous Random Variables

- X is continuous
- Can take any value
- How does one define probability?
  - $\sum_x P(X = x) = 1$
- Probability that X lies in an interval $[a, b]$?
  - $P(a < X \leq b) = P(x \leq b) - P(x \leq a)$
  - $F(q) = P(x \leq q)$ is the **cumulative distribution function**
  - $P(a < X \leq b) = F(b) - F(a)$

Probability Density

Probability Density Function

$$p(x) = \frac{\partial}{\partial x} F(x)$$

$$P(a < X \leq b) = \int_a^b p(x)dx$$

- Can $p(x)$ be greater than 1?

Expectation

- Expected value of a random variable

$$\mathbb{E}[X]$$

- What is most likely to happen in terms of X?
- For discrete variables

$$\mathbb{E}[X] \triangleq \sum_{x \in \mathcal{X}} xP(X = x)$$

- For continuous variables

$$\mathbb{E}[X] \triangleq \int_{\mathcal{X}} xp(x)dx$$

- **Mean** of X (μ)

Expectation of Functions of Random Variable

- Let $g(X)$ be a function of X
- If X is discrete:

$$\mathbb{E}[g(X)] \triangleq \sum_{x \in \mathcal{X}} g(x)P(X = x)$$

- If X is continuous:

$$\mathbb{E}[g(X)] \triangleq \int_{\mathcal{X}} g(x)p(x)dx$$

Properties

- $\mathbb{E}[c] = c$, c - constant
- $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$
- $\mathbb{E}[aX] = a\mathbb{E}[X]$
- Jensen's inequality: If $\varphi(X)$ is convex,

$$\varphi(\mathbb{E}[X]) \leq \mathbb{E}[\varphi(X)]$$

Variance

- Spread of the distribution

$$\begin{aligned} var[X] &\triangleq \mathbb{E}((X-\mu)^2) \\ &= \mathbb{E}(X^2) - \mu^2 \end{aligned}$$

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

What is a Probability Distribution?

Discrete

- Binomial, *Bernoulli*
- Multinomial, *Multinolli*
- Poisson
- Empirical

Continuous

- Gaussian (Normal)
- Degenerate pdf
- Laplace
- Gamma
- Beta
- Pareto

Binomial Distribution

- X = Number of heads observed in n coin tosses
- Parameters: n, θ
- $X \sim Bin(n, \theta)$
- Probability mass function (*pmf*)

$$Bin(k|n, \theta) \triangleq \binom{n}{k} \theta^k (1 - \theta)^{n-k}$$

Bernoulli Distribution

- Binomial distribution with $n = 1$
- Only one parameter ($\boldsymbol{\theta}$)

Multinomial Distribution

- Simulates tossing a K sided dice n times
- Random vector $\boldsymbol{x} = (x_1, x_2, \ldots, x_K)$
- Parameters: $n, \boldsymbol{\theta} \leftarrow \Re^K$, θ_j - probability that j^{th} side shows up
- $\boldsymbol{x} \sim Mu(n, \boldsymbol{\theta})$

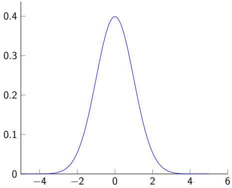
$$Mu(\mathbf{x}|n, \boldsymbol{\theta}) \triangleq \binom{n}{x_1, x_2, \ldots, x_K} \prod_{j=1}^{K} \theta_j^{x_j}$$

Multinoulli Distribution

- Multinomial distribution with $n = 1$
- $\mathbf{x}$ is a vector of 0s and 1s with only one bit set to 1, called **one-hot vector**
- Only one parameter (θ)

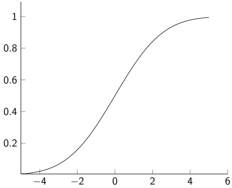
Gaussian (Normal) Distribution

$$\mathcal{N}(x|\mu,\sigma^2) \triangleq \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$



- Parameters:
  1. $\mu = \mathbb{E}[X]$
  2. $\sigma^2 = var[X] = \mathbb{E}[(X-\mu)^2]$
- $X \sim \mathcal{N}(0,1) \Leftarrow X$ is a **standard normal random variable**
- **Cumulative distribution function**:

$$\Phi(x;\mu,\sigma^2) \triangleq \int_{-\infty}^{x} \mathcal{N}(z|\mu,\sigma^2)dz$$



Joint Probability Distributions

- Multiple *related* random variables
- $p(x_1, x_2, \ldots, x_D)$ for $D > 1$ variables $(X_1, X_2, \ldots, X_D)$
- Discrete random variables: multi-dimensional array of size $O(K^D)$
- Continuous random variables: certain functional form

Covariance

- How does X vary with respect to Y
- For linear relationship:

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

Covariance and Correlation

- $\mathbf{x}$ is a d-dimensional random vector

$$cov[\mathbf{X}] \triangleq \mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^\top]$$

$$= \begin{pmatrix} var[X_1] & cov[X_1, X_2] & \cdots & cov[X_1, X_d] \\ cov[X_2, X_1] & var[X_2] & \cdots & cov[X_2, X_d] \\ \vdots & \vdots & \ddots & \vdots \\ cov[X_d, X_1] & cov[X_d, X_2] & \cdots & var[X_d] \end{pmatrix}$$

- Normalized covariance $\Rightarrow$ **Correlation**

Correlation

- *Pearson* Correlation Coefficient

$$corr[X, Y] \triangleq \frac{cov[X, Y]}{\sqrt{var[X]var[Y]}}$$

- What is $corr[X, X]$?
- $-1 \leq corr[X, Y] \leq 1$
- When is $corr[X, Y] = 1$?

Correlation

- *Pearson* Correlation Coefficient

$$corr[X, Y] \triangleq \frac{cov[X, Y]}{\sqrt{var[X]var[Y]}}$$

- What is $corr[X, X]$?
- $-1 \leq corr[X, Y] \leq 1$
- When is $corr[X, Y] = 1$?
  - $Y = aX + b$

Multivariate Gaussian Distribution

- Most widely used joint probability distribution

$$\mathcal{N}(\mathbf{X}|\mu, \mathbf{\Sigma}) \triangleq \frac{1}{(2\pi)^{D/2}|\mathbf{\Sigma}|^{1/2}} exp\left[-\frac{1}{2}(\mathbf{x}-\mu)^\top \mathbf{\Sigma}^{-\mathbf{1}}(\mathbf{x}-\mu)\right]$$

Linear Transformations of Random Variables

- What is the distribution of $f(\mathbf{X})$ ($\mathbf{X} \sim p()$)?
  - Linear transformation:

$Y = \mathbf{a}^\top \mathbf{X} + b$

- $\mathbb{E}[Y] = a^\top \mu + b$
- $var[Y] = \mathbf{a}^\top \mathbf{\Sigma} \mathbf{a}$

$\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{b}$

- $\mathbb{E}[Y] = \mathbf{A}\mu + \mathbf{b}$
- $cov(Y) = \mathbf{A}\mathbf{\Sigma}\mathbf{A}^\top$

- The Matrix Cookbook [2]
- `http://orion.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf`
- Available on Piazza

General Transformations

- $f()$ is **not linear**
- Example: X - discrete

$$Y = f(X) = \begin{cases} 1 & \text{if } X \text{ is even} \\ 0 & \text{if } X \text{ is odd} \end{cases}$$

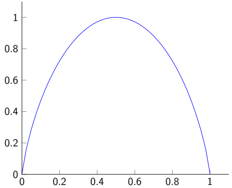
Monte Carlo Approximation

- Generate N samples from distribution for X
- For each sample, $x_i, i \in [1, N]$, compute $f(x_i)$
- Use empirical distribution as *approximate* true distribution

Approximate Expectation

$$\mathbb{E}[f(X)] = \int f(x)p(x)dx \approx \frac{1}{N}\sum_{i=1}^{N} f(x_i)$$

- Quantifying uncertainty of a random variable



Entropy

- $\mathbb{H}(X)$ or $\mathbb{H}(p)$

$$\mathbb{H}(X) \triangleq -\sum_{k=1}^{K} p(X=k)\log_2 p(X=k)$$

  - Variable with maximum entropy: Uniform distribution
  - Lowest entropy: Zero

Comparing Two Distributions

- **Kullback-Leibler Divergence** (or *KL Divergence* or *relative entropy*): Measuring the dissimilarity of two probability distributions

$$
\begin{aligned}
\mathbb{KL}(p||q) &\triangleq \sum_{k=1}^{K} p(k) \log \frac{p_k}{q_k} \\
&= \sum_k p(k) \log p(k) - \sum_k p(k) \log q(k) \\
&= -\mathbb{H}(p) + \mathbb{H}(p, q)
\end{aligned}
$$

- $\mathbb{H}(p, q)$ is the *cross-entropy*
- KL-divergence is asymmetric
- *Important fact:* $\mathbb{H}(p, q) \geq \mathbb{H}(p)$

Mutual Information

- What does learning about one variable X tell us about another, Y?
  - Correlation?

Mutual Information

$$\mathbb{I}(X;Y) = \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

- $\mathbb{I}(X;Y) = \mathbb{I}(Y;X)$
- $\mathbb{I}(X;Y) \geq 0$, equality iff $X \perp Y$

References

E. Jaynes and G. Bretthorst.
Probability Theory: The Logic of Science.
Cambridge University Press Cambridge:, 2003.

K. B. Petersen and M. S. Pedersen.
The matrix cookbook, nov 2012.
Version 20121115.

L. Wasserman.
All of Statistics: A Concise Course in Statistical Inference (Springer Texts in Statistics).
Springer, Oct. 2004.