CS 419M: Intro to ML

Overview Lecture: Overview of probability theory for ML

A review of probability theory

- Sample space(S): A sample space is defined as a set of all possible outcomes of an experiment. (For now *discrete*.)

  Example of an experiment: a coin pair toss; $S = \{HH, HT, TH, TT\}$.

- Probability Distribution p: A function $p : S \to [0, 1]$ s.t. $\sum_{x \in S} p(x) = 1$.

  E.g., $p(HH) = 1/3, p(HT) = 1/3, p(TH) = 1/6, p(TT) = 1/6$.

- Event (E) : An event is any subset of the sample space. i.e., $E \subseteq S$.

  Often we describe events using "conditions." E.g., the event that the two coin tosses are the same, $E_{\text{same}} = \{HH, TT\}$.
  Events can be combined using logical operations on these conditions: E.g., "the two coin tosses are the same **or** the second one is T":
  $E = \{HH, TT\} \cup \{HT, TT\} = \{HH, TT, HT\}$.

A review of probability theory

- Probability of an Event: The total weight assigned to all the elements in the event by a given probability distribution.

$$\Pr(E) = \sum_{x \in E} p(x)$$

- Note:
  - $\Pr(S) = 1$ and $\Pr(\emptyset) = 0$
  - $\Pr(\overline{E}) = 1 - \Pr(E)$, where $\overline{E} = S \setminus E$
  - $\Pr(E_1 \cup E_2) = \Pr(E_1) + \Pr(E_2) - \Pr(E_1 \cap E_2)$
  - If $E_1, E_2, \ldots, E_n$ are pairwise disjoint events, then

$$Pr(\bigcup_{i=1}^{n} E_i) = \sum_{i=1}^{n} Pr(E_i)$$

A review of probability theory

- Conditional Probability: For events E_1, E_2 (s.t. $\Pr(E_2) > 0$), define

$$\Pr(E_1|E_2) = \frac{\Pr(E_1 \cap E_2)}{\Pr(E_2)}$$

- Bayes' Rule, named after Thomas Bayes (1701-1761):

$$\Pr(A|B) = \frac{\Pr(B|A)\Pr(A)}{\Pr(B)},$$

provided $\Pr(A), \Pr(B) > 0$.

Exercise

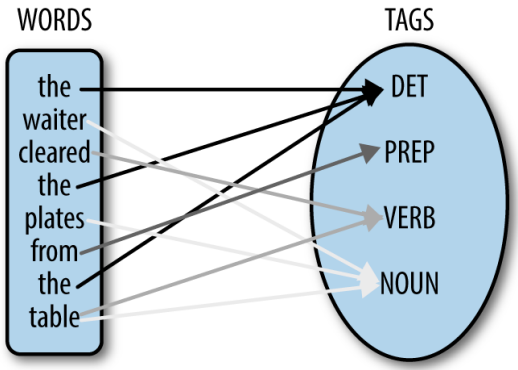
A lab test has a probability 0.95 of detecting a disease when applied to a person suffering from said disease, and a probability 0.10 of giving a false positive when applied to a non-sufferer. If 0.5% of the population are sufferers, what is the probability of a test being positive?

Example - Part of Speech

POS tagging is a popular Natural Language Processing (**NLP**) problem

Input: A set of n-words

Output: Part-of-speech (POS) tag for each word



Assuming each word is independently drawn from a fixed vocabulary, find the probability that a sentence of length m contains a 'noun' given that it contains a 'verb'.

Solution:

- Probability that a word is of POS type 'k' is p_k
- Let A_k be the event that the sentence contains POS type 'k'

$\Pr(A_k) = 1 - (1 - p_k)^m$
where $(1 - p_k)^m$ is that all m words are not of POS type 'k'.

$$\Pr(A_{noun} \mid A_{verb}) = \frac{\Pr(A_{noun} \cap A_{verb})}{\Pr(A_{verb})}$$

$$\begin{aligned} \Pr(A_j \cap A_k) &= 1 - \Pr(\overline{A_j \cap A_k}) \\ &= 1 - \Pr(\overline{A_j} \cup \overline{A_k}) \end{aligned} \tag{1}$$

We know that $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$. Now, we need $\Pr(\overline{A_j} \cap \overline{A_k})$. This refers to the event that neither tag k nor tag j are present. Thus, $\Pr(\overline{A_j} \cap \overline{A_k}) = (1 - (p_j + p_k))^m$. Using this in Equation 1, we get

$$\begin{aligned} \Pr(A_j \cap A_k) &= 1 - \Pr(\overline{A_j}) - \Pr(\overline{A_k}) + (1 - (p_j + p_k))^m \\ &= 1 - (1 - p_j)^m - (1 - p_k)^m + (1 - (p_j + p_k))^m \end{aligned}$$

Final answer:

$$Pr(A_{noun} \mid A_{verb}) = \frac{1-(1-p_{noun})^m-(1-p_{verb})^m+(1-p_{noun}-p_{verb})^m}{1-(1-p_{verb})^m}$$

Random Variable

- Random variable (X) : A variable that takes values from S according to a probability distribution.
  E.g., $S = \{H, T\}$ and $\Pr(X = H) = 2/3$, $\Pr(X = T) = 1/3$.
- Jointly distributed random variables: X, Y take values from sets S_1, S_2 respectively, according to a distribution over $S_1 \times S_2$. (Generalizes to more than 2 r.v.s.)
  E.g., $S_1 = S_2 = \{H, T\}$ and
  $\Pr((X, Y) = (H, H)) = 1/3$, $\Pr((X, Y) = (H, T)) = 1/3$,
  $\Pr((X, Y) = (T, T)) = 1/6$, $\Pr((X, Y) = (T, H)) = 1/6$.
- Marginal Distribution: For $x \in S_1$, $\Pr(X = x) = \sum_{y \in S_2} \Pr((X, Y) = (x, y))$. Similarly, marginal distribution of Y.

Independence of Random Variables

- Abuse of Notation: $\Pr((X, Y) = (x, y))$ abbreviated as $p(x, y)$, $\Pr(X = x | Y = y)$ as $p(x|y)$ etc.
- X, Y are said to be **independent** ($X \perp\!\!\!\perp Y$) iff for all x, y, the events $X = x$ and $Y = y$ are independent. i.e., $p(x, y) = p(x)p(y)$.
- X, Y are said to be **conditionally independent** given Z iff for all x, y, z (with $p(z) > 0$), $p(x, y|z) = p(x|z)p(y|z)$.

Expectation

If X is a random variable taking (say) real values (i.e., $S \subseteq \mathbb{R}$), we can define an "expected value" for X as:

$$E(X) = \Sigma_{x \in S} x \Pr(X = x)$$

Question: Suppose, $f : S \to \mathbb{R}$ and Y is the random variable jointly distributed with X defined as $Y = g(X)$. Then, what is $E(Y)$?

$$E(Y) = \sum_{x \in S} g(x) \Pr(X = x)$$

Properties of Expectation

1. $E[X + Y] = E[X] + E[Y]$ Proof HW

2. $E[(X - c)^2] \geq E[(X - \mu)^2]$ Proof HW
   where $\mu = E[X]$
   for any constant c and any random variable X

3. $E[cX] = cE[X]$ Proof HW
   for any constant c and any random variable X

4. If X, Y are independent, then $E[XY] = E[X]E[Y]$ Proof HW.

Variance

The **variance** of a random variable X with $E[X] = \mu$ is defined as:

$$Var[X] = E[(X - \mu)^2]$$

- $Var[X] = E[X^2] - (E[X])^2$
- $Var[X + \beta] = Var[X]$ and $Var[\alpha X] = \alpha^2 Var[X]$
- If $X_1, \cdots, X_n$ are *pairwise independent*, then $Var[\sum_i X_i] = \sum_i Var[X_i]$ (Proof HW)
- If $X_1, \cdots, X_n$ are pairwise independent, each with variance σ^2, then, $Var[\frac{1}{n} \sum_i X_i] = \sigma^2/n$

Chebyshev's Inequality

If variance is small, the random variable has to be (somewhat) *concentrated* around its mean.

If X is a random variable with mean μ and variance σ^2, then $\forall \alpha > 0$

$$Pr[|X - \mu| \geq \alpha] \leq \frac{\sigma^2}{\alpha^2}$$

If X_i are pairwise independent, identically distributed random variables with mean μ and variance σ^2, then $\Pr[|\left(\frac{1}{n}\sum_i X_i\right) - \mu| \geq \alpha] \to 0$ as $n \to \infty$.
Law of Large Numbers: Sample mean $\frac{1}{n}\sum_i X_i$ (for say, i.i.d. samples) converges to the expected value μ as n grows.

Recall: Expectation, Variance

If X is a random variable, taking (say) real values (i.e., $S \subseteq \mathbb{R}$), with probability distribution $p : S \to [0, 1]$, then **expected value** of X is:

$$E[X] = \sum_{x \in S} p(x) \cdot x$$

The **variance** of a random variable X with $E[X] = \mu$ is

$$Var[X] = E[(X - \mu)^2] = E[X^2] - \mu^2$$

Covariance

For random variables X and Y, **covariance** is defined as:

$$Cov[X, Y] = E[(X - E(X))(Y - E(Y))] = E[XY] - E[X]E[Y]$$

If X and Y are independent then their covariance is 0, since in that case

$$E[XY] = E[X]E[Y]$$

Note: However, covariance being 0 does not necessarily imply that the variables are independent.

Covariance

Properties:

1. $Cov[X, X] = Var[X]$
2. $Cov[X + Z, Y] = Cov[X, Y] + Cov[Z, Y]$
3. $Cov[\sum_i X_i, Y] = \sum_i Cov[X_i, Y]$

Some Important Distributions

Notation: We write $X \sim \mathcal{D}$ if the random variable X takes its values according to some distribution $\mathcal{D} : S \to [0, 1]$: i.e., $\Pr[X = x] = \mathcal{D}(x)$.

Bernoulli Random Variable: Takes values from $S = \{0, 1\}$.

A single **parameter** $q \in [0, 1]$ fully specifies a Bernoulli random variable, where $\Pr[X = 1] = q$ (and $Pr[X = 0] = 1 - q$).
Notation: Denoted as $\mathrm{Bern}(q)$.

If $X \sim \mathrm{Bern}(q)$ then

- $E[X] = (1 - q) \times 0 + q \times 1 = q$
- $Var[X] = q - q^2 = q(1 - q)$

Quintessential example: $X = 1$ iff a coin toss yields H, where q is the coin's probability of coming up H.

Binomial Random Variable

Quintessential example: number of H out of n tosses of a coin with $\Pr[H] = q$.

Sample space, $S = \{0, \cdots, n\}$.

Notation: Distribution denoted as $\mathrm{B}(n, q)$. Note: **2 parameters**, n and q.

Formally, if $Y = \sum_{i=1}^{n} X_i$, where $X_i \sim \mathrm{Bern}(q)$ are i.i.d., then $Y \sim \mathrm{B}(n, q)$.

If $Y \sim \mathrm{B}(n, q)$ then

- $\Pr(Y = k) = \binom{n}{k} q^k (1-q)^{n-k}$
- $E[Y] = nq$ (by linearity of expectation, applied to $Y = \sum_{i=1}^{n} X_i$)
- $Var[Y] = nq(1-q)$ (Why?)

Coming Up

Continuous Random Variables

- Distribution function
  - Probability Density Function
  - Cumulative Distribution Function
  - Conditioning, Independence for Continuous Random Variables
  - Expectation (and other statistics) for Continuous Random Variables
  - Important Example: Normal Distribution
- Central Limit Theorem

Continuous Distributions

Suppose a random variable X is uniformly distributed in the range $S = [0, 1]$, and Y is uniformly distributed in $[0, \frac{1}{2}]$. But note that for all $a \in [0, 1]$, $\Pr(X = a) = \Pr(Y = a) = 0$! So, how do we formally define these distributions?

Mass vs. Density: When S is discrete, weights were assigned to individual items in S so that they summed up to 1. When S is continuous, we shall assign *density* to all points in S so that S still "weighs" 1.

A **probability density function** (pdf) of a continuous random variable is $f : S \to \mathbb{R}_+$ such that for all $D \subseteq S$, $Pr(X \in D) = \int_D f(x)dx$.

E.g., Above, $f_X(a) = 1$ for all $a \in [0, 1]$, but $f_Y(a) = \begin{cases} 2 & \text{if } a \leq \frac{1}{2} \\ 0 & \text{otherwise} \end{cases}$

Cumulative Distribution Function

Suppose X is a continuous random variable which takes values from the sample space $\mathbb{R}$, and has a pdf f. Its **cdf** is defined as $F : \mathbb{R} \to [0, 1]$:

$$F(a) = Pr(X \leq a) = \int_{-\infty}^{a} f(x)dx$$

Note: pdf for continuous distribution can be obtained by differentiating the cdf of that random variable:

$$f(a) = \frac{dF(x)}{dx}|_{x=a}$$

Joint Distributions

A jointly distributed pair of continuous random variables X, Y has a joint distribution function $f : \mathbb{R}^2 \to \mathbb{R}_+$ s.t. for all $D \subseteq \mathbb{R}^2$,

$$\Pr((X, Y) \in D) = \int \int_D f(x, y) dxdy.$$

If $f(x, y)$ is a joint pdf, then

$$F(a, b) = Pr(X \leq a, Y \leq b) = \int_{-\infty}^{b} \int_{-\infty}^{a} f(x, y) dxdy$$

$$f(a, b) = \frac{\partial^2 F(x, y)}{\partial x \partial y}|_{a,b}$$

Independence

Suppose X, Y are two continuous random variables with a joint pdf f.

Marginal Distribution: From f can define marginal distribution density functions f_X and f_Y: $f_X(a) = \int_{-\infty}^{\infty} f(a, y)dy$ and $f_Y(a) = \int_{-\infty}^{\infty} f(x, a)dx$

Independence: X, Y are said to be independent iff for all x, y,

$$f(x, y) = f_X(x)f_Y(y)$$

where f_X, f_Y are the marginal density functions.

Conditional Density

Suppose X and Y are two jointly distributed continuous random variables, with a density function f. Then we can define the conditional probability density of X given $Y = y$ as:

$$f_X(x|Y = y) = \frac{f(x, y)}{f_Y(y)} = \frac{f(x, y)}{\int_{-\infty}^{\infty} f(x, y)dx}$$

If X, Y are independent, then for all x, y (with non-zero marginals), $f_X(x|Y = y) = f_X(x)$ and $f_Y(y|X = x) = f_Y(y)$.

Example

Let X and Y be *independent continuous* random variables with same marginal density functions

$$f(x) = \begin{cases} e^{-x} & \text{if } x > 0; \\ 0 & \text{otherwise.} \end{cases}$$

Define the random variable $Z = \frac{X}{Y}$. Find the density function for Z.

Solution

$$F_Z(a) = 0 \qquad \text{if } a \leq 0$$

$$\begin{aligned} F_Z(a) &= \Pr(Z \leq a) = \Pr(X \leq aY) \\ &= \int_0^\infty \int_0^{ya} f(x,y)dxdy = \int_0^\infty \int_0^{ya} e^{-x}e^{-y}dxdy \\ &= \int_0^\infty e^{-y}(1-e^{-ya})dy \\ &= 1 - \frac{1}{a+1} = \frac{a}{a+1} \qquad \text{if } a > 0 \end{aligned}$$

The pdf is obtained by differentiating the cdf. For $a > 0$:

$$f_Z(a) = \frac{dF_Z(z)}{dz}|_a = \frac{d}{dz}(1 - \frac{1}{z+1})|_a = \frac{1}{(a+1)^2}.$$

Expectation

Expectation is equivalent to probability density weighted integral of possible values.

$$E[X] = \int_{-\infty}^{\infty} xf(x)dx$$

Note: Expectation of a function of a random variable is

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)dx$$

Variance, Covariance are defined as before in terms of expectation

Normal (Gaussian) Distribution

- It is a popular continuous distribution

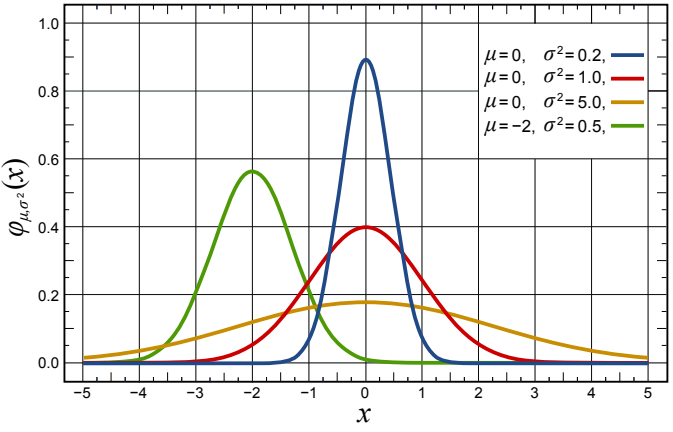
$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

  - μ is the mean and σ^2 is the variance

- Exercise: Verify the mean and variance. For e.g.

$$E[X] = \int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) dx \stackrel{?}{=} \mu$$

1-D Gaussian distribution



Normal (Gaussian) Distribution

- It is a popular continuous distribution

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

  - μ is the mean and σ^2 is the variance

- Exercise: Verify the mean and variance. For e.g.

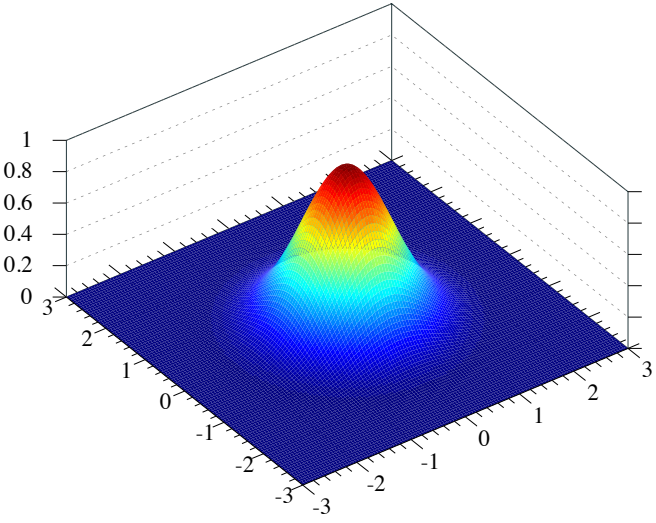
$$E[X] = \int_{-\infty}^{\infty} x\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)dx \stackrel{?}{=} \mu$$

- Multivariate Gaussian (d-dim)

$$f(x|\mu,\Sigma) = (2\pi)^{-d/2}|\Sigma|^{-1/2}\exp\left\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\}$$

  - x is now a vector, μ is the mean vector and Σ is the co-variance matrix

2-D Gaussian distribution



Properties of Normal Distribution

- All marginals of a Gaussian are again Gaussian
- Any conditional of a Gaussian is Gaussian
- The product of two Gaussians is again Gaussian
- Even the sum of two independent Gaussian r.v.'s is a Gaussian

Note: Many of the standard distributions belong to the family of **exponential distributions**

- Bernoulli, binomial/multinomial, Poisson, Normal (Gaussian), Beta/Dirichlet ...
- Share many important properties - e.g. They have a conjugate prior. (We will briefly discuss this within the next 2 weeks)

Central Limit Theorem

Suppose $E[X] = \mu$ and $X_1, X_2, \ldots$ are independent r.v.s distributed as X

Define $S^{(m)} = \frac{1}{m} \sum_{i=1}^{m} X_i$ (average of m independent copies of X)

Law of large numbers: for very large m, $S^{(m)}$ is sharply concentrated around μ

But for moderately large m, $S^{(m)}$ is a little "fuzzy" around μ. What does this distribution look like?

Depends on the distribution of X. In particular, $S^{(1)}$ is identical to X.

But as m increases, it very quickly starts looking like a Gaussian distribution, *no matter what the actual distribution of X is*![1]

[1] Actually, the rate of convergence depends on $E[|X - \mu|^3]$.

Central Limit Theorem: More Formally

Suppose $X_1, X_2, \ldots$ are i.i.d random variables, each with mean μ and variance σ^2. Let $S^{(m)} = \frac{1}{m}\sum_{i=1}^{m} X_i$.

Note $E[S^{(m)}] = \mu$ and $Var[S^{(m)}] = \sigma^2/m$

To observe the shape of the distribution without letting it becoming sharper as m grows, we shall scale it to make variance a constant: Let $Z^{(m)} = \sqrt{m}(S^{(m)} - \mu)$, so that $E[Z^{(m)}] = 0$ and $Var[Z^{(m)}] = \sigma^2$.

Central Limit Theorem: As $m \to \infty$, the distribution of $Z^{(m)} \to \mathcal{N}(0, \sigma^2)$.

Rule of Thumb: For $m > 30$, we may take $Z^{(m)} \sim \mathcal{N}(0, \sigma^2)$. i.e., $S^{(m)} \sim \mathcal{N}(\mu, \sigma^2/m)$.