



COMPSCI 389

Introduction to Machine Learning

Probability, Statistics, and Quantifying Uncertainty

Prof. Philip S. Thomas (pthomas@cs.umass.edu)

Random Variable

- A random variable is a mathematical formulation of a quantity that depends on random events.
- We use upper case letters to represent random variables (e.g., X) and lower-case to represent constants (e.g., x).
- We can talk about the probability of a random variable X taking a value x : $\Pr(X = x)$.
- Example:
 - If X is a roll of a fair die, then $\Pr(X = 3) = 1/6$.
- A full characterization of random variables is beyond the scope of this course, and can be a surprisingly deep topic (see “measure theoretic probability”).

Probability Distribution

- A probability distribution (probability measure) gives the probability that a random variable takes different values.
 - Technically it gives the probability of events (not necessarily values or outcomes), but a formal characterization of “events” is beyond the scope of this class.
- We can talk about the “distribution of a random variable.”
- Example:
 - Let p be the distribution of a fair die.
 - $p(1) = p(2) = p(3) = p(4) = p(5) = p(6) = \frac{1}{6}$
 - For all such discrete distributions: $\forall x, p(x) \geq 0$ and $\sum_x p(x) = 1$.

Probability Distribution (continued)

- We often say that we have multiple random variables “sampled from the same distribution”.
- Here “sampled” is slightly imprecise.
- We really mean that we have multiple random variables, they all have the same distribution, and they are all statistically independent.
 - **i.i.d.:** Independent and identically distributed.
- Example:
 - Let X_1 and X_2 be two random variables, each representing a sample of a fair die.
 - If the two die rolls are independent, what is $\Pr(X_1 + X_2 = 7)$?

Realization or Instance of a Random Variable

- Once a random variable has been sampled, it takes a specific value.
- This is called a *realization* or *instance* of the random variable.
- A realization of a random variable is a **constant**.
- Let x_1 and x_2 denote the realization of two fair die rolls.
- What is $\Pr(x_1 = x_2)$?
 - Trick question! There is nothing random here. They are either equal or not, and so this probability is either 0 or 1.
 - Think of x_1 and x_2 as symbols in place of specific numbers.
 - What is $\Pr(3 = 3)$? What is $\Pr(1 = 2)$?

Random Data Sets

- In ML, we typically think of data sets as being random samples from some distribution, called the **data generating distribution**.
 - **Example:** The GPA data set contains samples from the distribution of students applying to UFRGS.
- We may write (X, Y) to denote a random variable representing one sample from this distribution.
- A data set contains many of these random variables: $(X_i, Y_i)_{i=1}^n$.
 - This data set is itself a random quantity!
 - We can reason about things like $\Pr(X_1 = X_2)$, $\Pr(Y_1 = Y_2 | X_1 \neq X_2)$, or even the probability of the MSE of the model learned by NN being below a constant value!

Random Data Sets: Example

- Consider a data set containing $n = 2$ rolls of a fair die.
- X_1 and X_2 are random variables representing independent rolls of the die:

$$\Pr(X_1 = 1) = \Pr(X_1 = 2) = \Pr(X_1 = 3) = \Pr(X_1 = 4) = \Pr(X_1 = 5) = \Pr(X_1 = 6) = \frac{1}{6}$$

- The data set is (X_1, X_2) .
- What is $\Pr(X_1 = X_2)$?

Non-Random Data Sets

- The data set that we see is one sample of the random variables.
- Once we have the data set as a computer file, it is no longer random, and so we write: $(x_i, y_i)_{i=1}^n$.
- In the die example, the data set is (x_1, x_2) .
- Here x_1 and x_2 are symbols representing numbers (not random!).
- What is $\Pr(x_1 = x_2)$?
 - It's either zero or one! Either they are equal or not. There is nothing random about $x_1 = x_2$!

Random vs Non-Random

- Note: Different ML texts take different random/not-random perspectives for data sets!
 - Texts emphasizing principled theory typically take the random perspective.
 - Texts emphasizing basic practice typically take the non-random perspective.
- When writing pseudocode for an algorithm, should we view the data as random or non-random?
 - No agreed-upon convention!

Random vs Non-Random Terminology

- The terms *random* and *non-random* are imprecise.
 - People often use random to mean “uniform random.”
 - Its precise meaning is “is a random variable.”
 - A random variable can always take the same value, effectively being constant!
- Random → Stochastic (avoids confusion with “uniform random”)
- Non-Random → Deterministic or constant (cannot be “random”).

Probability and Statistics Terminology

- **Parameter / Population Statistic:** A parameter is a property of a probability distribution (or random variable), like the mean or variance.
 - Example: Mean $\mathbf{E}[X]$
- **Sample:** One or more “draws” of a random variable.
 - $X_1, X_2, \dots, X_n$ might be random variables representing n samples.
 - Example: These represent n rolls of the same die
 - Often samples $X_1, X_2, \dots, X_n$ are *independent and identically distributed* (i.i.d.).
 - $x_1, x_2, \dots, x_n$ might be the realization of n samples.
 - Example: The actual outcomes of n rolls of a die.
 - It is not meaningful to discuss whether $x_1, x_2, \dots, x_n$ are i.i.d.

Probability and Statistics Terminology

- **Statistic / Sample Statistic:** Statistics are properties of a sample. To emphasize this, we sometimes say “sample statistic.”
 - Example: Sample mean $\frac{1}{n} \sum_{i=1}^n X_i$
 - Notice that the sample mean is itself a random variable!
 - We can also consider a realization of the sample mean: $\frac{1}{n} \sum_{i=1}^n x_i$.

Mean Squared Error (revisited)

- The MSE is:

$$\text{MSE} = \mathbf{E} \left[(Y - \hat{Y})^2 \right].$$

- This is a *parameter* or *population statistic*.
- The sample MSE is:

$$\widehat{\text{MSE}}_n = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \quad \text{or} \quad \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2.$$

- This is a *statistic* or *sample statistic*.
- The “hat” means “an estimate” and the n -subscript indicates it is computed from n samples.
- Our goal is typically to optimize a parameter.
 - We don’t know this parameter’s value.
- In an attempt to achieve this goal, we use sample statistics.
 - We can compute sample statistics from data!

Can we trust sample statistics?

- How much we should trust sample statistics depends on:
 - The number of samples, n .
 - If the average of 3 die rolls is 4, and the average of 3,000 die rolls is 3.47, which do you trust more?
 - The variance of the samples.
 - Consider the samples $(-1, -0.3, 0, 0.5, 0.8)$ versus $(-820, -214, 12, 480, 542)$
 - Both have sample mean 0. Which are you more confident has a mean in the range $[-10, 10]$?
- Idea: Use the number of samples and variance of samples to estimate how accurate the sample statistic is.

Confidence Interval

- We will use the number of samples and their variance to construct a **confidence interval** for the parameter (e.g., MSE) based on the sample statistic (sample MSE).
- A confidence interval is an interval (range of numbers) that contains a parameter with a specified confidence, $1 - \delta$.
- If $[L, U]$ is a $1 - \delta$ confidence interval for the mean μ , then
$$\Pr(L \leq \mu \leq U) \geq 1 - \delta.$$
- **Question:** What is random in this statement of probability?
- **Answer:** The *confidence interval* is random! It is typically computed from data. Different samples of data result in different lower and upper bounds.

Standard Error

- One common way to obtain a confidence interval is using **standard error**.
- Let $x_1, x_2, \dots, x_n$ be a sequence of n numbers.
- Let σ be the sample **standard deviation** of this sequence (with [Bessel's correction](#)):

$$\sigma = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}},$$
$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

- The **standard error** is then

$$SE = \frac{\sigma}{\sqrt{n}}.$$

Using Standard Error

- If $X_1, X_2, \dots, X_n$ are n random variables and:
 - The random variables are i.i.d. with mean μ .
 - The random variables are each normally distributed.
 - $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ is the sample mean.
- Then $[\bar{X} - 1.96 \times SE, \bar{X} + 1.96 \times SE]$ is a 95% confidence interval for μ .
- That is:

$$\Pr(\bar{X} - 1.96 \times SE \leq \mu \leq \bar{X} + 1.96 \times SE) \geq 0.95.$$

- **Note:** There exist other confidence intervals for the mean that don't assume that data is normal (e.g., Maurer & Pontil), and even confidence intervals that don't assume independence (e.g., Azuma) or identically distributed samples (e.g., Hoeffding)!
 - In general, all confidence intervals make *some* assumptions, but the assumptions differ.
 - Confidence intervals with weaker assumptions tend to be “loose” (have wide intervals).

“The random variables are each normally distributed”

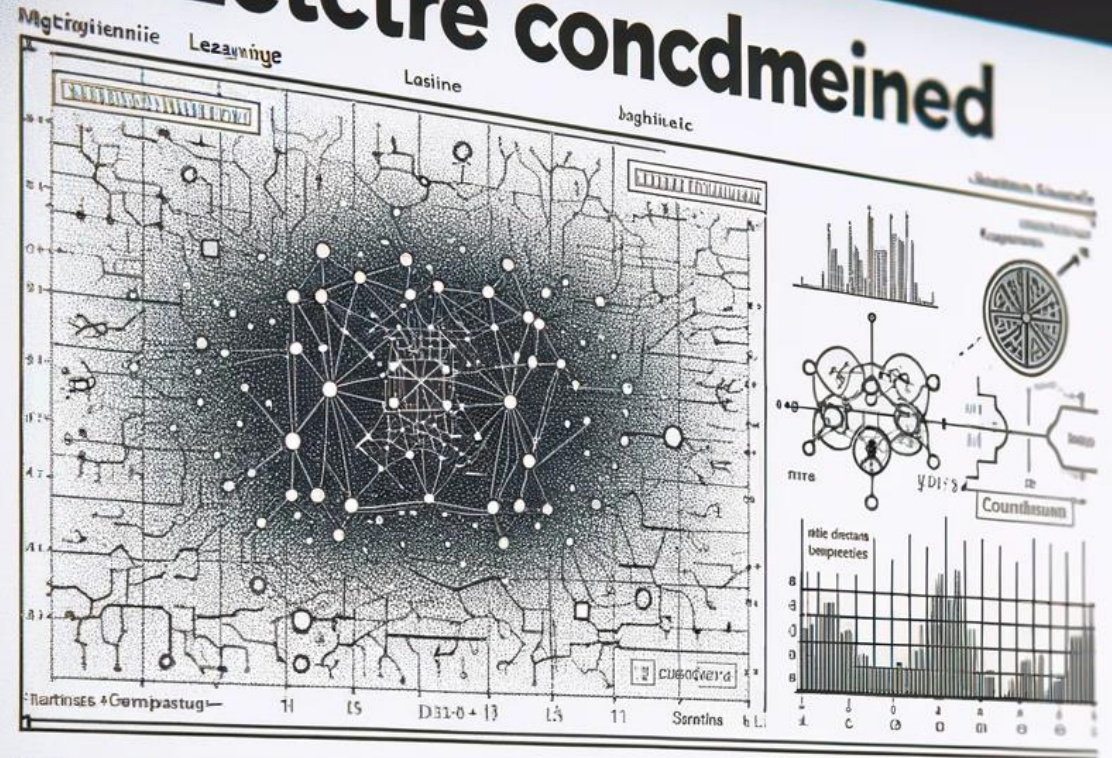
- Actually, we only require the sample mean to be normally distributed.
- **Question:** Why is it reasonable to assume the sample mean is normally distributed?
- **Answer:** Central Limit Theorem
 - As $n \rightarrow \infty$, the sample mean becomes normally distributed regardless of the sampling distribution.
 - So, when n is “big enough”, this assumption is “reasonable” (still false though...)
 - What value of n is “big enough” depends on the problem.
 - I’ve seen examples where 20 is enough and where hundreds of thousands are not enough.

Mean Squared Error (re-revisited)

- MSE: $\text{MSE} = \mathbf{E} \left[(Y - \hat{Y}_i)^2 \right]$.
- Sample MSE: $\widehat{\text{MSE}}_n = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$.
- Let $Z_i = (Y_i - \hat{Y}_i)^2$.
- Notice that $\mu = \mathbf{E}[Z_i] = \text{MSE}$, and let SE be the standard error of $Z_1, Z_2, \dots, Z_n$.
- So, $\widehat{\text{MSE}}_n \pm 1.96 \times \text{SE}$ is a 95% confidence interval for the actual MSE (under normality assumptions).
 - Although normality assumptions often false, this gives a rough idea of how much the sample MSE can be trusted.

End

Letctre concdmeined



Dgainmnic



Mbcime Learning

Thank you.

