CMPT 210: Probability and Computing

Lecture 24

Sharan Vaswani

April 6, 2023

Recap – Machine Learning 101

Q: Suppose we toss a coin (with unknown bias) 10 times and get the sequence $HHTTHTTTTH$. What is the probability that I will see $4H$ and $6T$ in the next 10 tosses of the coin?

Collect data: $\mathcal{D} = \{H, H, T, T, H, T, T, T, T, H\}$ when tossing the coin 10 times.

Assume model: Each toss is independent and follows the same Bernoulli distribution.

Construct the likelihood function: $\Pr[\mathcal{D}|p] = p^4 (1-p)^6$

Compute the MLE by training the model: The MLE $\hat{p}$ is our estimate of Pr[heads] that maximizes the likelihood of seeing the data. Hence,

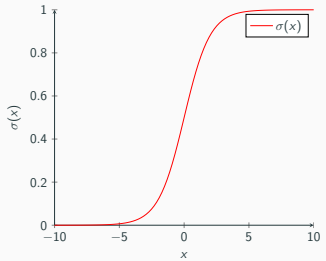
$$\hat{p} = \arg\max_p \Pr[\mathcal{D}|p] = \arg\min_p -\log(\Pr[\mathcal{D}|p]) = \arg\min_p [4\log(p) + 6\log(1-p)]$$

$$\implies \frac{4}{\hat{p}} - \frac{6}{1-p} = 0 \implies \hat{p} = \frac{4}{10}. \qquad \text{(Our estimate of Pr[heads] is } \hat{p} = 0.4\text{)}$$

Prediction: Since we estimate that each coin flip follows the same Bernoulli distribution with Pr[heads] $= 0.4$, $\Pr[\text{observe 4H, 6T}|\hat{p} = 2/5] = \binom{10}{4} (0.4)^4 (1-0.4)^6$

Digression – Sigmoid function

The **sigmoid** function is defined as: $\sigma : \mathbb{R} \to [0, 1]$: $\sigma(x) := \frac{1}{1+\exp(-x)}$.



Since the range of σ is $[0, 1]$, we will use it to output probabilities. In particular, define **parameter** $\theta \in \mathbb{R}$ s.t. $p = \sigma(\theta) = \frac{1}{1+\exp(-\theta)}$.

σ is an invertible function and hence there is a one-one mapping from θ to p (every p can be specified by specifying the equivalent θ).

Note that $1 - p = 1 - \frac{1}{1+\exp(-\theta)} = \frac{\exp(-\theta)}{1+\exp(-\theta)} = \frac{1}{1+\exp(\theta)}$. Hence, if $p = \sigma(\theta)$, $1 - p = \sigma(-\theta)$.

Machine Learning 101 – Estimating the bias of multiple coins

Q: Suppose we toss 5 different coins and obtain the sequence $HTHHT$. We want to estimate the bias of each of these coins, but have additional information that the bias of coin i depends on its weight $x_i \in \mathbb{R}$ (which is known).

Collect data: $\mathcal{D} = \{H, T, H, H, T\}$ when tossing 5 different coins.

Assume model: The bias of each coin depends on its weight according to a *linear model* i.e.

$$p_i = \sigma(\theta x_i) = \frac{1}{1 + \exp(-\theta x_i)}$$

Here, θ is the **parameter** of our model, x_i (the known weights) for the coins are referred to as the **features**. The model is **linear** because the argument to the sigmoid function is linear in θ.

Construct the likelihood function: Given θ, the coin tosses are independent, i.e. if p_i is the bias of coin i, the likelihood in terms of θ is given by:

$$\begin{aligned}\Pr[\mathcal{D}|\{x_1, x_2, \ldots, x_5\}, \theta] &= p_1\,(1 - p_2)\,p_3\,p_4\,(1 - p_5) \\ &= [\sigma(\theta x_1)]\,[\sigma(-\theta x_2)]\,[\sigma(\theta x_3)][\sigma(\theta x_4)][\sigma(-\theta x_5)]\end{aligned}$$

Machine Learning 101 – Estimating the bias of multiple coins

Simplify the likelihood function: Define $y_i \in \mathbb{R}$ such that $y_i = 1$ if coin i in $\mathcal{D}$ is a heads and $y_i = -1$ if coin i in $\mathcal{D}$ is a tails. For $\mathcal{D} = HTHHT$, $y_1 = 1$, $y_2 = -1$ and so on. For example i, y_i is referred to as the **label**, hence each toss i is described by the (x_i, y_i) pair referred to as the **input-output** pair or the **feature-label** pair.

$$\Pr[\mathcal{D}|\{x_1, x_2, \ldots, x_5\}, \theta] = [\sigma(y_1\theta x_1)]\,[\sigma(y_2\theta x_2)]\,[\sigma(y_5\theta x_5)] = \prod_{i=1}^{5} [\sigma(y_i\theta x_i)]$$

$$\implies -\log(\Pr[\mathcal{D}|\{x_1, x_2, \ldots, x_5\}, \theta]) = \sum_{i=1}^{5} -\log(\sigma(y_i\theta x_i)) = \sum_{i=1}^{5} \log\left(1 + \exp(-y_i\theta x_i)\right)$$

The NLL defined above is a function of θ and is referred to as the **logistic loss**. This model is referred to as (1-dimensional) **logistic regression**. Since we are classifying the coins as those that came up heads or tails, we are doing **binary classification**.

Logistic regression for binary classification is heavily used in machine learning. E.g. Classifying whether a patient with features such as their medical history, symptoms, test results has cancer.

Machine Learning 101 – Estimating the bias of multiple coins

Compute the MLE by training the model: In order to compute the MLE $\hat{\theta}$, we need to minimize the NLL on the previous slide,

$$\hat{\theta} = \arg\min_{\theta} \sum_{i=1}^{5} \log\left(1 + \exp(-y_i \theta x_i)\right).$$

Unfortunately, this optimization problem can not be solved directly by taking derivatives like before. We need techniques from **numerical optimization** that studies efficiently minimizing complicated functions and the related computational properties (for example, see `https://vaswanis.github.io/409_981-F22.html`).

Prediction: Once we have computed $\hat{\theta}$, we can use it to predict the probability of heads for a new coin that has feature x as $\hat{p} = \sigma(\hat{\theta} x)$.

Machine Learning 101 – Generalizing to multiple dimensions

Suppose each coin i has a vector of features – its weight, color resistance, etc that affect its bias. If there are d such features, $x_i \in R^d$.

Assume model: We will have a vector of parameters $\theta \in \mathbb{R}^d$ and the bias of each coin depends on its features as:

$$p_i = \sigma\left(\sum_{j=1}^{d} \theta_j \, x_{i,j}\right) \qquad (x_{i,j} \text{ corresponds to feature } j \text{ of coin } i)$$

Writing this in terms of the dot product $\langle \theta, x_i \rangle = \sum_{j=1}^{d} \theta_j x_{i,j}$,

$$p_i = \frac{1}{1 + \exp(-\langle \theta, x_i \rangle)}$$

Construct the likelihood function: Suppose we toss the 5 coins, and get the same $\mathcal{D} = \{H, T, H, H, T\}$ sequence. By following the same steps as before, the MLE can be given by:

$$\hat{\theta} = \arg\min_{\theta} \sum_{i=1}^{5} \log\left(1 + \exp(-y_i \langle \theta x_i \rangle)\right).$$

Machine Learning 101 – Generalizing beyond coins

Suppose we are given n inputs in the form of their features (X) and labels y. Here, X is an $d \times n$-dimensional matrix such that the feature of input i is column $X_i \in \mathbb{R}^d$ and y is a n-dimensional vector such that $y_i \in \{-1, 1\}$.

In our coin example, the inputs were coins with different properties (features such as weight, color) and the labels corresponded to whether we got a heads ($y = 1$) or tails ($y = -1$).

The feature-label could be characteristics of a patient and whether or not they have cancer, pixels in pictures of cats and dogs and whether it is a cat or a dog, text in the email and whether or not it is spam (Gmail uses a logistic regression model for classifying spam).

Using the same linear model $\Pr[y_i = 1] = p_i = \frac{1}{1+\exp(-\langle\theta, X_i\rangle)}$ and constructing the likelihood function in the same manner,

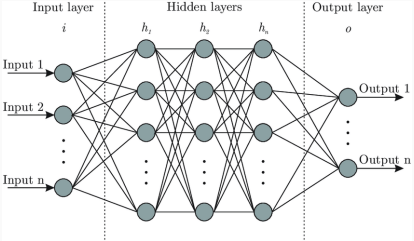
$$\hat{\theta} = \arg\min_{\theta} \underbrace{\sum_{i=1}^{n} \log\left(1 + \exp(-y_i \langle \theta X_i \rangle)\right)}_{\text{Logistic regression loss function}}$$

This is the definition you will find in machine learning textbooks!

Machine Learning 101 – Generalizing beyond linear models

We can choose alternative ways to model p_i. We have been using a linear model such that, $p_i = \sigma(\langle \theta, X_i \rangle)$, but we could use *any* function $f(\theta, X_i)$ as an argument to the sigmoid function.

Designing such f is a major research direction. Current most popular models (used to classify videos on YouTube, rank posts on Facebook/Instagram) are (much) larger variants of **neural networks** that look like this:



If you found this lecture this fascinating, you can take the CMPT 410 (Machine Learning) course offered in the Fall.

Questions?

Wrapping up

Sample (outcome) space $\mathcal{S}$: Nonempty (countable) set of possible outcomes.

Outcome $\omega \in \mathcal{S}$: Possible "thing" that can happen.

Event E: Any subset of the sample space.

Probability function on a sample space $\mathcal{S}$ is a total function $\Pr : \mathcal{S} \to [0, 1]$. For any $\omega \in \mathcal{S}$,

$$0 \leq Pr[\omega] \leq 1 \quad ; \quad \sum_{\omega \in \mathcal{S}} \Pr[\omega] = 1 \quad ; \quad \Pr[E] = \sum_{\omega \in E} \Pr[\omega]$$

Wrapping up

Union: For mutually exclusive events $E_1, E_2, \ldots, E_n$,
$\Pr[E_1 \cup E_2 \cup \ldots E_n] = \Pr[E_1] + \Pr[E_2] + \ldots + \Pr[E_n]$.

Complement rule: $\Pr[E] = 1 - \Pr[E^c]$

Inclusion-Exclusion rule: For any two events E, F, $\Pr[E \cup F] = \Pr[E] + \Pr[F] - \Pr[E \cap F]$.

Union Bound: For any events $E_1, E_2, E_3, \ldots E_n$, $\Pr[E_1 \cup E_2 \cup E_3 \ldots \cup E_n] \leq \sum_{i=1}^{n} \Pr[E_i]$.

Uniform probability space: A probability space is said to be uniform if $\Pr[\omega]$ is the same for every outcome $\omega \in \mathcal{S}$. In this case, $\Pr[E] = \frac{|E|}{|\mathcal{S}|}$.

Wrapping up

Conditional Probability: For events E and F, probability of event E conditioned on F is given by $\Pr[E|F]$ and can be computed as $\Pr[E|F] = \frac{\Pr[E \cap F]}{\Pr[F]}$.

Probability rules with conditioning: For the complement E^c, $\Pr[E^c|F] = 1 - \Pr[E|F]$.

Conditional Probability for multiple events:
$\Pr[E_1 \cap E_2 \cap E_3] = \Pr[E_1] \Pr[E_2|E_1] \Pr[E_3|E_1 \cap E_2]$.

Bayes rule: For events E and F if $\Pr[E] \neq 0$, $\Pr[F|E] = \frac{\Pr[E|F] \Pr[F]}{\Pr[E]}$.

Law of Total Probability: For events E and F, $\Pr[E] = \Pr[E|F] \Pr[F] + \Pr[E|F^c] \Pr[F^c]$.

Independent Events: Events E and F are said to be independent, if knowledge that F has occurred does not change the probability that E occurs, i.e. $\Pr[E|F] = \Pr[E]$ and $\Pr[E \cap F] = \Pr[E] \Pr[F]$.

Wrapping up

Pairwise Independence: Events E_1, E_2, ..., E_n are pairwise independent, if for *every* pair of events E_i and E_j ($i \neq j$), $\Pr[E_i|E_j] = \Pr[E_i]$ and $\Pr[E_i \cap E_j] = \Pr[E_i]\ \Pr[E_j]$.

Mutual Independence: Events E_1, E_2, ..., E_n are mutually independent, if for *every* subset of events, the probability that all the selected events occur equals the product of the probabilities of the selected events. Formally, for every subset $S \subseteq \{1, 2, \ldots, n\}$, $\Pr[\cap_{i \in S} E_i] = \prod_{i \in S} \Pr[E_i]$.

Random variable: A random "variable" R on a probability space is a total function whose domain is the sample space $\mathcal{S}$. The codomain is denoted by V (usually a subset of the real numbers), meaning that $R : \mathcal{S} \to V$.

Indicator Random Variables: An indicator random variable corresponding to an event E is denoted as $\mathcal{I}_E$ and is defined such that for $\omega \in E$, $\mathcal{I}_E[\omega] = 1$ and for $\omega \notin E$, $\mathcal{I}_E[\omega] = 0$.

Wrapping up

Probability density function (PDF): Let R be a random variable with codomain V. The probability density function of R is the function $\mathrm{PDF}_R : V \to [0,1]$, such that $\mathrm{PDF}_R[x] = \Pr[R = x]$ if $x \in \mathrm{Range(R)}$ and equal to zero if $x \notin \mathrm{Range(R)}$.

$\sum_{x \in V} \mathrm{PDF}_R[x] = \sum_{x \in \mathrm{Range(R)}} \Pr[R = x] = 1$.

Cumulative distribution function (CDF): The cumulative distribution function of R is the function $\mathrm{CDF}_R : \mathbb{R} \to [0,1]$, such that $\mathrm{CDF}_R[x] = \Pr[R \leq x]$.

Distribution over a random variable can be fully specified using the cumulative distribution function (CDF) (usually denoted by F). The corresponding probability density function (PDF) is denoted by f.

Wrapping up

Bernoulli Distribution: $f_p(0) = 1 - p$, $f_p(1) = p$. *Example*: When tossing a coin such that Pr[heads] $= p$, random variable R is equal to 1 if we get a heads (and equal to 0 otherwise). In this case, R follows the Bernoulli distribution i.e. $R \sim \text{Ber}(p)$.

Uniform Distribution: If $R : \mathcal{S} \to V$, then for all $v \in V$, $f(v) = 1/|V|$. *Example*: When throwing an n-sided die, random variable R is the number that comes up on the die. $V = \{1, 2, \ldots, n\}$. In this case, R follows the Uniform distribution i.e. $R \sim \text{Uniform}\{1, 2, \ldots, n\}$.

Binomial Distribution: $f_{n,p}(k) = \binom{n}{k} p^k (1-p)^{n-k}$. *Example*: When tossing n independent coins such that Pr[heads] $= p$, random variable R is the number of heads in n coin tosses. In this case, R follows the Binomial distribution i.e. $R \sim \text{Bin}(n, p)$.

Geometric Distribution: $f_p(k) = (1-p)^{k-1} p$. *Example*: When repeatedly tossing a coin such that Pr[heads] $= p$, random variable R is the number of tosses needed to get the first heads. In this case, R follows the Geometric distribution i.e. $R \sim \text{Geo}(p)$.

Wrapping up

Expectation/mean of a random variable R is denoted by $\mathbb{E}[R]$ and "summarizes" its distribution. Formally, $\mathbb{E}[R] := \sum_{\omega \in \mathcal{S}} \Pr[\omega]\, R[\omega]$

Alternate definition of expectation: $\mathbb{E}[R] = \sum_{x \in \text{Range}(R)} x \Pr[R = x]$.

Expectation of transformed r.v's: For a random variable $X : \mathcal{S} \to V$ and a function $g : V \to \mathbb{R}$, we define $\mathbb{E}[g(X)]$ as follows: $\mathbb{E}[g(X)] := \sum_{x \in \text{Range}(X)} g(x) \Pr[X = x]$

Linearity of Expectation: For n random variables $R_1, R_2, \ldots, R_n$ and constants $a_1, a_2, \ldots, a_n$, $b_1, b_2, \ldots, b_n$, $\mathbb{E}\left[\sum_{i=1}^{n} a_i R_i + b_i\right] = \sum_{i=1}^{n} a_i \mathbb{E}[R_i] + b_i$.

Conditional Expectation: For random variable R, the expected value of R conditioned on an event A is given by $\mathbb{E}[R|A] = \sum_{x \in \text{Range}(R)} x \Pr[R = x|A]$

Law of Total Expectation: If R is a random variable $\mathcal{S} \to V$ and events $A_1, A_2, \ldots A_n$ form a partition of the sample space, then, $\mathbb{E}[R] = \sum_i \mathbb{E}[R|A_i] \Pr[A_i]$.

Wrapping up

Independent random variables: We define two random variables R_1 and R_2 to be independent if for *all* $x_1 \in \text{Range}(R_1)$ and $x_2 \in \text{Range}(R_2)$, events $[R_1 = x_1]$ and $[R_2 = x_2]$ are independent. More formally,

$$\Pr[(R_1 = x_1) \cap (R_2 = x_2)] = \Pr[(R_1 = x_1)] \Pr[(R_2 = x_2)]$$

Independent random variables: Two random variables R_1 and R_2 are independent if for *all* $x_1 \in \text{Range}(R_1)$ and $x_2 \in \text{Range}(R_2)$,

$$\Pr[(R_1 = x_1) | (R_2 = x_2)] = \Pr[(R_1 = x_1)]$$
$$\Pr[(R_2 = x_2) | (R_1 = x_1)] = \Pr[(R_2 = x_2)]$$

Expectation of product of r.v's: For two r.v's R_1 and R_2, $\mathbb{E}[R_1 R_2] = \sum_{x \in \text{Range}(R_1 R_2)} x \Pr[R_1 R_2 = x]$.

Expectation of product of independent r.v's: For independent r.v's R_1 and R_2, $\mathbb{E}[R_1 R_2] = \mathbb{E}[R_1] \mathbb{E}[R_2]$.

Wrapping up

Joint distribution between r.v's X and Y can be specified by its joint PDF as follows: $\text{PDF}_{X,Y}[x, y] = \Pr[X = x \cap Y = y]$.

If X and Y are independent random variables, $\text{PDF}_{X,Y}[x, y] = \text{PDF}_X[x] \, \text{PDF}_Y[y]$.

Marginalization: We can obtain the distribution for each r.v. from the joint distribution by marginalizing over the other r.v's i.e. $\text{PDF}_X[x] = \sum_i \text{PDF}_{X,Y}[x, y_i]$.

Variance: Standard way to measure the deviation from the mean. For r.v. X, $\text{Var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2] = \sum_{x \in \text{Range}(X)} (x - \mu)^2 \Pr[X = x]$ where $\mu := \mathbb{E}[X]$.

Alternate definition of variance: $\text{Var}[X] = \mathbb{E}[X^2] - \mu^2 = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$.

Standard Deviation: For r.v. X, the standard deviation of X is defined as $\sigma_X := \sqrt{\text{Var}[X]} = \sqrt{\mathbb{E}[X^2] - (\mathbb{E}[X])^2}$.

Wrapping up

Properties of variance: For constants a, b and r.v. R, $\text{Var}[aR + b] = a^2\text{Var}[R]$.

Pairwise Independence of r.v's: Random variables $R_1, R_2, R_3, \ldots R_n$ are pairwise independent if for any pair R_i and R_j, for $x \in \text{Range}(R_i)$ and $y \in \text{Range}(R_j)$,
$\Pr[(R_i = x) \cap (R_j = y)] = \Pr[R_i = x] \Pr[R_j = y]$.

Linearity of variance for pairwise independent r.v's: If $R_1, \ldots, R_n$ are pairwise independent, $\text{Var}[R_1 + R_2 + \ldots R_n] = \sum_{i=1}^{n} \text{Var}[R_i]$.

Properties of variance: If $R_1, \ldots, R_n$ are pairwise independent, for constants $a_1, a_2, \ldots a_n$ and $b_1, b_2, \ldots b_n$, $\text{Var}[\sum_{i=1}^{n} a_i R_i + b_i] = \sum_{i=1}^{n} a_i^2 \text{Var}[R_i]$.

Wrapping up

Covariance: For two random variables R and S, the covariance between R and S is defined as: $\text{Cov}[R, S] = \mathbb{E}[(R - \mathbb{E}[R])(S - \mathbb{E}[S])] = \mathbb{E}[RS] - \mathbb{E}[R]\,\mathbb{E}[S]$.

Properties of covariance: If R and S are independent r.v's, $\mathbb{E}[RS] = \mathbb{E}[R]\,\mathbb{E}[S]$ and $\text{Cov}[R, S] = 0$. $\text{Cov}[R, R] = \text{Var}[R]$. $\text{Cov}[R, S] = \text{Cov}[S, R]$.

Variance of sum of r.v's: For r.v's $R_1, R_2, \ldots, R_n$,
$\text{Var}\left[\sum_{i=1}^n R_i\right] = \sum_{i=1}^n \text{Var}[R_i] + 2\sum_{1 \le i < j \le n} \text{Cov}[R_i, R_j]$.

If R_i and R_j are pairwise independent, $\text{Cov}[R_i, R_j] = 0$ and $\text{Var}\left[\sum_{i=1}^n R_i\right] = \sum_{i=1}^n \text{Var}[R_i]$.

Correlation: For two r.v's R_1 and R_2, the correlation between R_1 and R_2 is defined as $\text{Corr}[R_1, R_2] = \frac{\text{Cov}[R_1, R_2]}{\sqrt{\text{Var}[R_1]\,\text{Var}[R_2]}}$. $\text{Corr}[R_1, R_2] \in [-1, 1]$ and indicates the strength of the relationship between R_1 and R_2.

Wrapping up

Bernoulli: If $R \sim \text{Bernoulli}(p)$, $\mathbb{E}[R] = p$ and $\text{Var}[R] = p\,(1-p)$.

Uniform: If $R \sim \text{Uniform}(\{v_1, \ldots, v_n\})$, $\mathbb{E}[R] = \frac{v_1+v_2+\ldots+v_n}{n}$ and $\text{Var}[R] = \frac{[v_1^2+v_2^2+\ldots v_n^2]}{n} - \left(\frac{[v_1+v_2+\ldots v_n]}{n}\right)^2$.

Binomial: If $R \sim \text{Bin}(n, p)$, $\mathbb{E}[R] = np$ and $\text{Var}[R] = n\,p\,(1-p)$.

Geometric: If $R \sim \text{Geo}(p)$, $\mathbb{E}[R] = \frac{1}{p}$ and $\text{Var}[R] = \frac{1-p}{p^2}$.

Wrapping up

Tail inequalities bound the probability that the r.v. takes a value much different from its mean.

Markov's Theorem: If X is a non-negative random variable, then for all $x > 0$, $\Pr[X \geq x] \leq \frac{\mathbb{E}[X]}{x}$.

Chebyshev's Theorem: For a r.v. X and all $x > 0$, $\Pr[|X - \mathbb{E}[X]| \geq x] \leq \frac{\mathrm{Var}[X]}{x^2}$.

Weak Law of Large Numbers: Let $G_1, G_2, \ldots, G_n$ be pairwise independent variables with the same mean μ and (finite) standard deviation σ. Define $T_n := \frac{\sum_{i=1}^{n} G_i}{n}$, then for every $\epsilon > 0$, $\lim_{n\to\infty} \Pr[|T_n - \mu| \leq \epsilon] = 1$.

Chernoff Bound: If $T_1, T_2, \ldots, T_n$ are mutually independent r.v's such that $0 \leq T_i \leq 1$ for all i. If $T := \sum_{i=1}^{n} T_i$, for all $c \geq 1$ and $\beta(c) := c\ln(c) - c + 1$, $\Pr[T \geq c\mathbb{E}[T]] \leq \exp(-\beta(c)\,\mathbb{E}[T])$.

What's Next – Continuous Random Variables

We have studied random variables that can take on discrete values – number of heads when tossing a coin, the number on a dice or the number of attempts to hit the bullsye in a dart game.

We have used these discrete distributions for designing randomized algorithms for verifying matrix multiplication, finding the maximum cut in graphs, polling and binary classification in machine learning.

In many applications, it is often more natural to model quantities as continuous random variables, for example, the amount of time it takes to transmit a message over a noisy channel or study the distribution of income in a population.

Continuous random variables are often used in distributed computing and for machine learning – fitting a model that can effectively explain the collected data.

What's Next – Continuous Random Variables

Discrete random variables can take on specific values in an interval. For example, if $X \sim \text{Uniform}\{v_1, v_2, \ldots, v_n\}$, X can take on values from the set $\{v_1, v_2, \ldots, v_n\}$. If $X \sim \text{Bin}(n, p)$, X can take on values in the set $\{0, , 1, \ldots, n\}$.

Continuous random variable: A r.v. that can take on all possible values in a specified interval.

Example: If R is a continuous r.v. and $R \sim \text{Uniform}[0, 1]$, the r.v. R can be equal to any number in the $[0, 1]$ interval – for example, 0.01, $2/3$ or 0.9 with equal probability.

Continuous r.v. can often be seen as limits of discrete r.v. For example, the continuous uniform distribution is a limit of the discrete distribution $X \sim \text{Uniform}(v_1, v_2, \ldots, v_n)$ as $n \to \infty$.

Similarly, the normal (Gaussian) distribution can be interpreted as a limit of the Binomial distribution $X \sim \text{Bin}(n, p)$ as $n \to \infty$.

What's Next

STAT 271: Probability and Statistics for Computing Science

- Continuous random variables and distributions
- Sampling and Parameter estimation
- Linear Regression
- Hypothesis testing
- Analysis of Variance

Questions?