EECE 5644: Probability Theory

Mark Zolotas

E-mail: m.zolotas@northeastern.edu
Webpage: https://coe.northeastern.edu/people/zolotas-mark/

Tentative Course Outline (Wks. 1-2)

Topics	Dates	Assignments	Additional Reading
~~Course Overview~~ ~~Machine Learning Basics~~	~~07/05~~	**Optional Homework 0** released on Canvas on 07/08 but please do NOT submit on Canvas	~~Chpt. 1~~ ~~Murphy 2012~~
Foundations: ~~Linear Algebra~~, Probability, Numerical Optimization (Gradient Descent), Regression	07/06-11		~~Stanford LA Review~~ Stanford Prob. Review Chpt. 8 Murphy 2022
Quick Python Tutorial	07/12	**Homework 1** released on Canvas on 07/15 **Due 07/25**	N/A
Linear Classifier Design, Linear Discriminant Analysis and Principal Component Analysis (PCA)	07/13-14		Chpts. 9.2 & 20.1 Murphy 2022
Bayesian Decision Theory: Empirical Risk Min, Max Likelihood (ML), Max a Posteriori	07/14-15		Chpt. 2 Duda & Hart 2001 Deniz Erdogmus Notes

Linear Algebra Recap

- Inner product:

$$\mathbf{x} \cdot \mathbf{y} = \sum_{i=1}^{n} x_i y_i = \mathbf{x}^\intercal \mathbf{y} = \mathbf{y}^\intercal \mathbf{x}$$

- Eigenvalues/vectors for symmetric, square $\mathbf{A} = \mathbf{A}^\intercal \in \mathbb{R}^{n \times n}$

$$\mathbf{A}\mathbf{u}_i = \lambda_i \mathbf{u}_i \quad \text{for } i \in 1, \dots, n$$

- In matrix form:

"diagonilizable" *orthogonality*

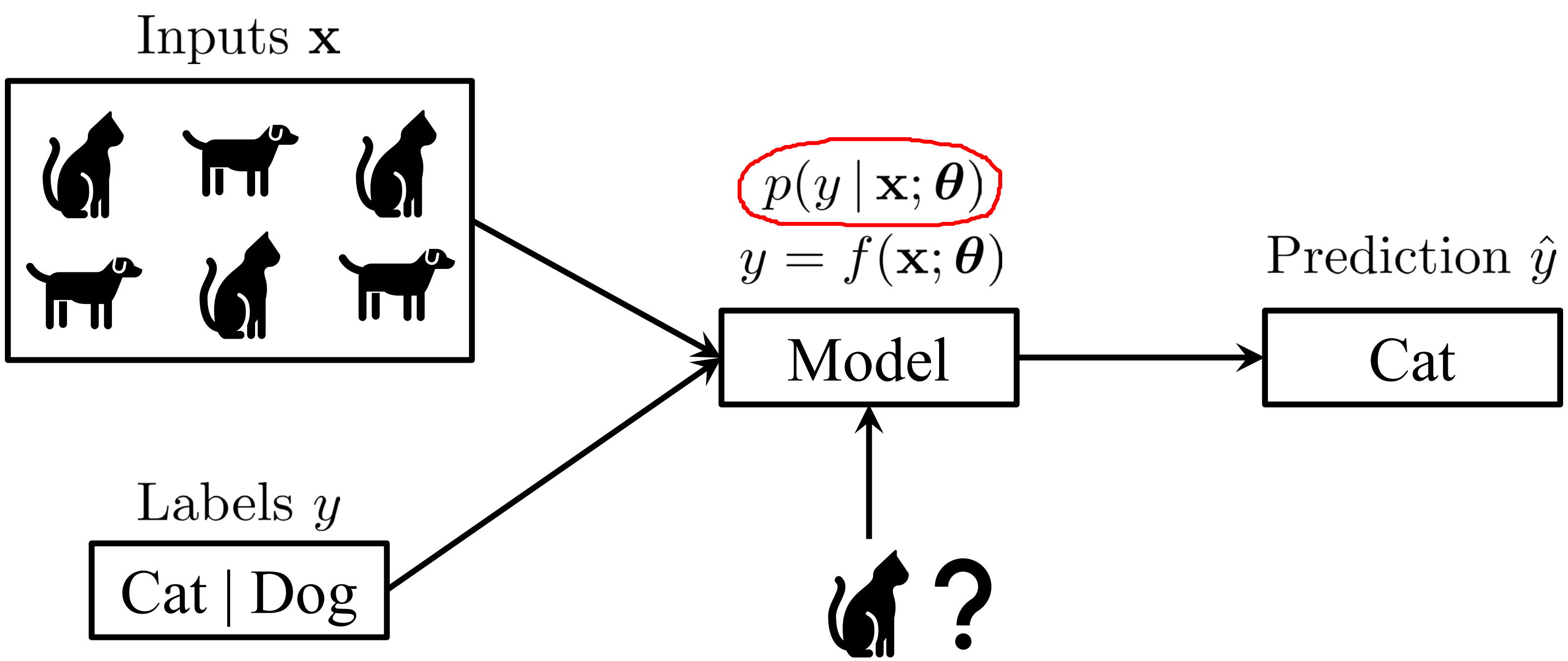
$$\mathbf{A}\mathbf{U} = \mathbf{U}\mathbf{\Lambda} \longrightarrow \mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{-1} \longrightarrow \mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\intercal$$

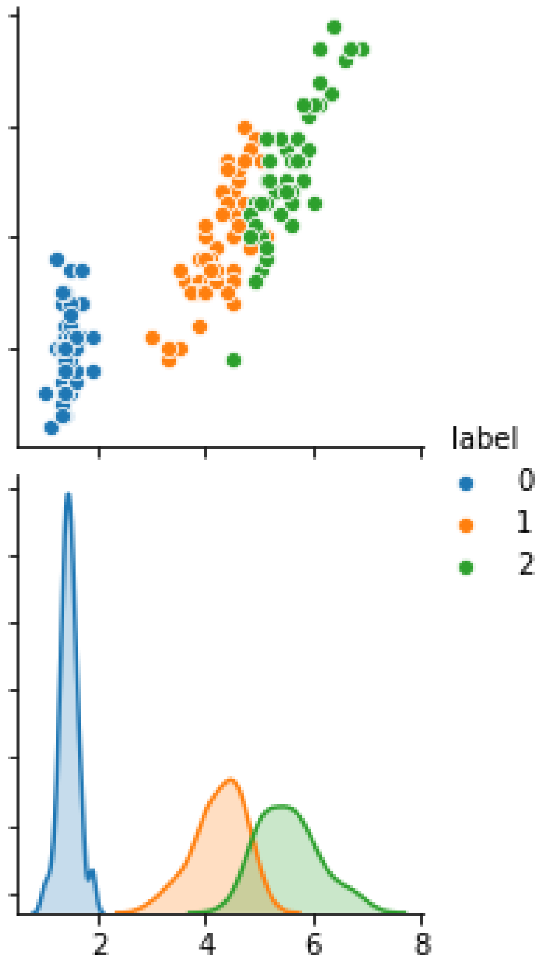
$\text{if } \mathbf{U}^{-1} \text{ exists}$ $\quad$ $\text{if } \mathbf{U}^\intercal\mathbf{U} = \mathbf{I}$

- Positive definiteness (PD) for $\mathbf{A} = \mathbf{A}^\intercal \in \mathbb{R}^{n \times n}$

$$\mathbf{A} > 0 \text{ iff } \mathbf{x}^\intercal \mathbf{A} \mathbf{x} > 0, \forall \mathbf{x} \in \mathbb{R}^n \setminus \{\mathbf{0}\} \text{ OR iff } \lambda_i > 0 \ \forall i$$

Probability Theory





Two Perspectives on Probability

- **Frequentist:** Concerned with repeated events and the *frequency* with which we expect to observe data, given some hypothesis about the world
  - Data treated as *random*, repeated trials might generate different data
  - Model parameters take a *single value* ("point estimate")
  - Parameters typically estimated by *maximum likelihood* of data
- **Bayesian:** Interested in the plausibility or uncertainty of a hypothesis, given evidence of data and our prior beliefs
  - Data treated as *fixed*, can make inferences about one-off events
  - Model parameters are *random variables* that have a probability distribution
  - Parameters estimated from data and *prior knowledge*
- **Model parameters** = Configuration variables learned from the data

Axioms of Probability

- Define an **event** A as a binary variable that holds or does not (true/false)
  - E.g. "it will be sunny tomorrow", "I have a headache", "I rolled a 6 in dice"
  - Each event has a probability $\Pr(A)$ of being true

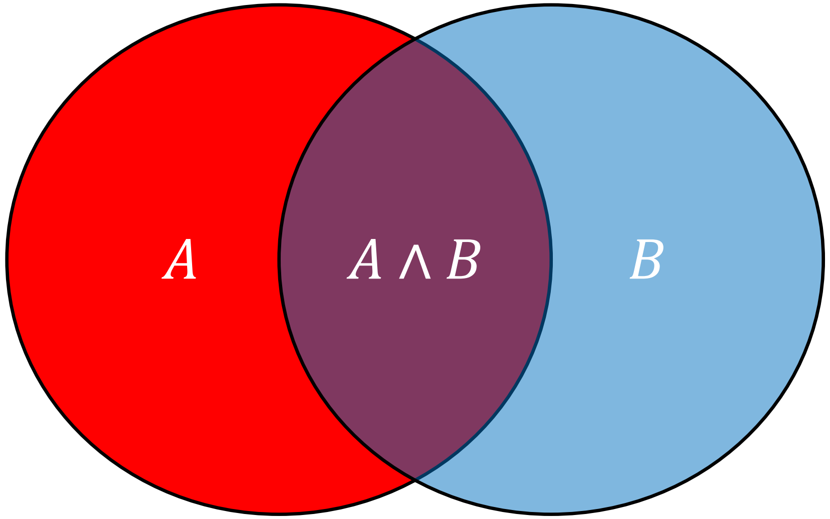
- Behind probability theory are 3 foundational axioms (Kolmogorov):
  1. All probabilities must satisfy $0 \leq \Pr(A) \leq 1$
  2. Valid event propositions (tautologies) have $\Pr(A) = 1$ and unsatisfiable facts (contradictions) have $\Pr(A) = 0$
  3. The union (disjunction) of two events is given by:

$$\Pr(A \vee B) = \Pr(A) + \Pr(B) - \Pr(A \wedge B)$$

If mutually exclusive

Conditional Probability

- Union/Disjunction: $\Pr(A \vee B) = \Pr(A) + \Pr(B) - \Pr(A \wedge B)$
- Joint Probability: $\Pr(A \wedge B) = \Pr(A, B) = \Pr(A)\Pr(B)$ If independent
- Conditional Probability: $\Pr(A|B) = \dfrac{\Pr(A, B)}{\Pr(B)}$



Random Variable (RV)

Def. Is a real-valued function, $X : \mathcal{X} \to \mathbb{R}$, that can take on values defined by a set of all possible outcomes, $\mathcal{X}$, known as the **sample space**. An **event** is a set of random outcomes from this sample space.

Example: If X is the result of a die rolled, then $\mathcal{X} = \{1, 2, ..., 6\}$, and the event of "rolling a 1" is denoted as $X = 1$, the event of "rolling even" is $X \in \{2, 4, 6\}$, the event of "rolling between 3 and 5" is $3 \leq X \leq 5$, etc.

Discrete Random Variables

- If sample space $\mathcal{X}$ is a **finite** number of distinct values, then X is **discrete**
- Probability of the event that X takes on value x is denoted as $\Pr(X = x)$
- Directly express this probability using a **probability mass function (PMF)**

$$p_X(x) = \Pr(X = x)$$

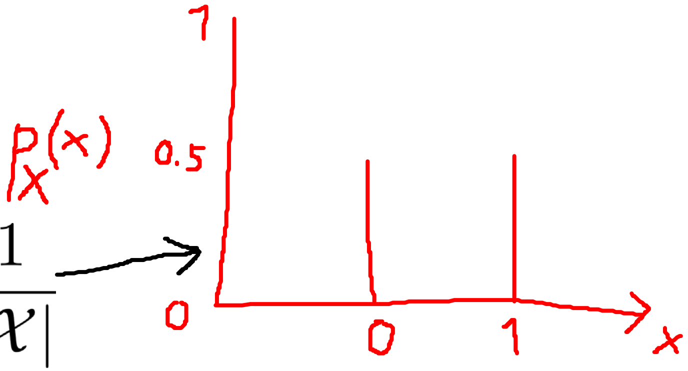
$$0 \leq p_X(x) \leq 1$$

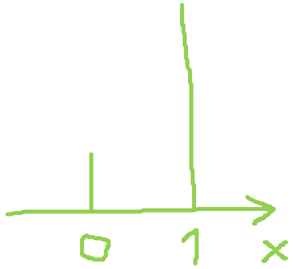
$$\sum_{x \in \mathcal{X}} p_X(x) = 1$$

- *Example:* X models a coin toss heads (1) or tails (0)

Uniform RV

$$p_X(x) = \frac{1}{|\mathcal{X}|}$$





$$p_X(x) = \begin{cases} 1 - p & \text{if } x = 0 \\ p & \text{if } x = 1 \end{cases}$$

Bernoulli RV

$X \sim \text{Ber}(p)$

Examples

- What about **multiple** random events?
- *Example:* X models number of heads in n coin tosses, what is the probability of k heads?

$$\Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Binomial RV

$X \sim \mathrm{Bin}(n, p)$

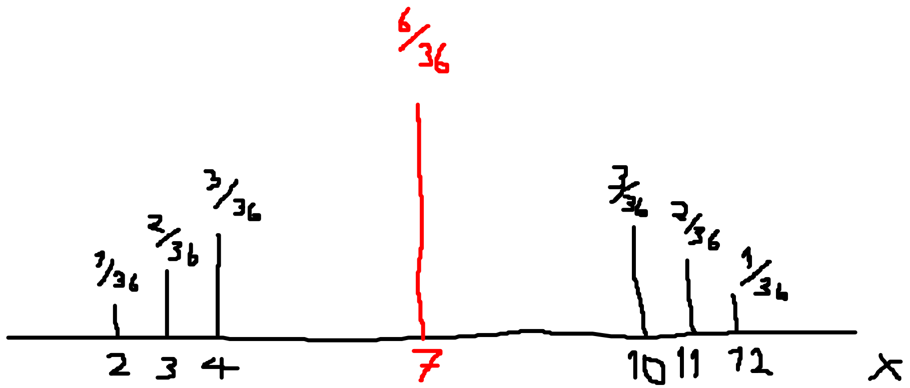
- *Example:* X models sum of two fair dice, what is $p_X(x)$ for $X \in \{2, \ldots 12\}$?

$$\Pr(X = 2) = \frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$$

$$\Pr(X = 4) = \frac{3}{36}$$

$$\Pr(X = 11) = \frac{2}{36}$$

What is highest $p_X(x)$?



Multiple Random Variables

- Let X and Y be discrete RVs, then the **joint distribution** is:

$$p_{XY}(x,y) = \Pr(X = x, Y = y) \qquad \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p_{XY}(x,y) = 1$$

- Define **marginal distribution** for X:

Sum/Total Probability Rule

$$p_X(x) = \sum_{y \in \mathcal{Y}} p_{XY}(x,y)$$

Process of summing out other RV known as "marginalization"

- Define **conditional distribution**:

Product Rule

Distribution over Y given that $X = x$

$$p_{Y|X}(y|x) = \frac{p_{X,Y}(x,y)}{p_X(x)} \iff p_{X,Y}(x,y) = p_{Y|X}(y|x)p_X(x)$$

Chain Rule of Probability

- Generalize product rule to n variables:

$$
\begin{aligned}
p_{X_1,\ldots,X_n}(\mathbf{x}_{1:n}) &= p(x_1, x_2, \ldots, x_n) \\
&= p(\mathbf{x}_{2:n} | x_1) p(x_1) \\
&= p(\mathbf{x}_{3:n} | x_1, x_2) p(x_2 | x_1) p(x_1) \\
&= p(x_n | \mathbf{x}_{1:n-1}) \ldots p(x_3 | x_1, x_2) p(x_2 | x_1) p(x_1)
\end{aligned}
$$

PMF subscript notation simplified in remainder of expression

Repeatedly apply rule of conditional probability

- Break down joint distribution into factorized form of conditionals until marginal in isolation; useful in machine learning

Conditional Independence

- Reminder of **unconditional** independence relation:

$$X \perp\!\!\!\perp Y \iff p_{X,Y}(x,y) = p_X(x)p_Y(y)$$

- Generalized to n variables: $$p_{X_1,\ldots,X_n}(\mathbf{x}_{1:n}) = \prod_i^n p_{X_i}(x_i)$$

- Rely more frequently on **conditional independence (CI)** between RVs:

$$\boxed{X \perp\!\!\!\perp Y \mid Z \iff p(x,y|z) = p(x|z)p(y|z)}$$

- *Example:* Look at 3rd term from left of chain rule $p(x_3|x_1,x_2)$

$$p(x_3|x_2,x_1) = \frac{p(x_3,x_2|x_1)}{p(x_2|x_1)} = \frac{p(x_3|x_1)p(x_2|x_1)}{p(x_2|x_1)} \qquad x_3 \perp\!\!\!\perp x_2 \mid x_1$$

Discrete RV Examples

- $X \sim Bernoulli(p)$ (where $0 \leq p \leq 1$): one if a coin with heads probability p comes up heads, zero otherwise.

$$p(x) = \begin{cases} p & \text{if } p = 1 \\ 1 - p & \text{if } p = 0 \end{cases}$$

- $X \sim Binomial(n, p)$ (where $0 \leq p \leq 1$): the number of heads in n independent flips of a coin with heads probability p.

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

- $X \sim Geometric(p)$ (where $p > 0$): the number of flips of a coin with heads probability p until the first heads.

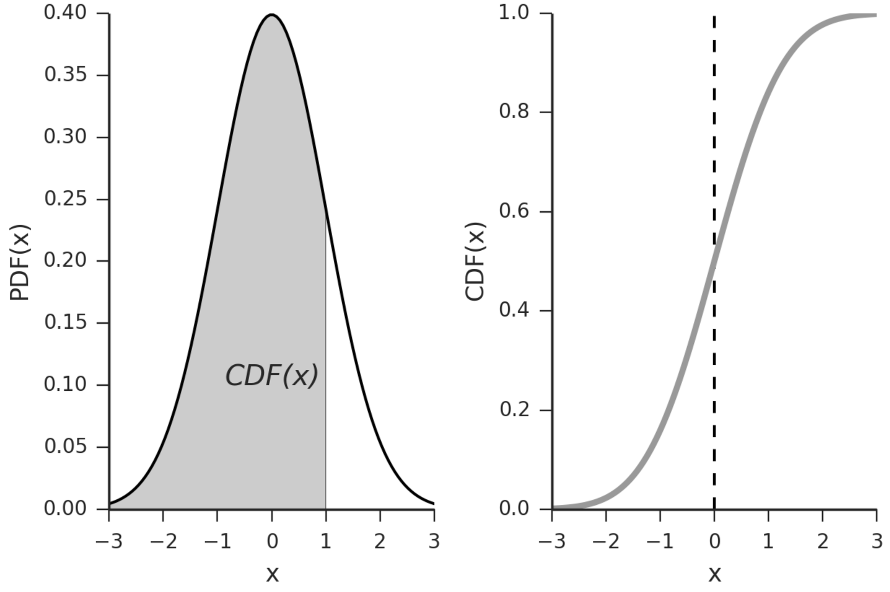
$$p(x) = p(1-p)^{x-1}$$

- $X \sim Poisson(\lambda)$ (where $\lambda > 0$): a probability distribution over the nonnegative integers used for modeling the frequency of rare events.

$$p(x) = e^{-\lambda} \frac{\lambda^x}{x!}$$

Continuous Random Variables

- If sample space $\mathcal{X}$ is **NOT countable**, then $X \in \mathbb{R}$ is **continuous**
- Can count *intervals* along this real line
- Define **cumulative density function (CDF)** as: $P_X(x) = \Pr(X \leq x)$
- **Probability density function (PDF)** as derivative: $p_X(x) = \frac{d}{dx} P_X(x)$



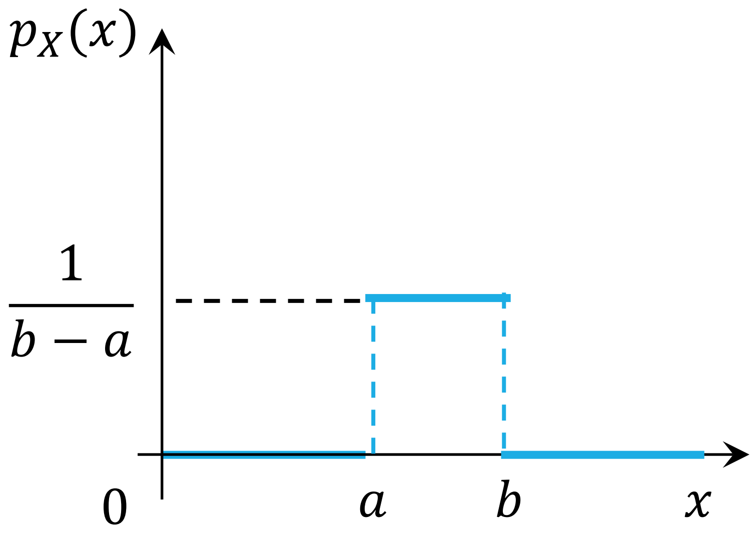
Note:

- CDF *non-decreasing* $\Rightarrow p_X(x) \geq 0$
- If CDF not differentiable, neither exist
- $p_X(x) \neq \Pr(X = x)$, possible for $p_X(x) > 1$

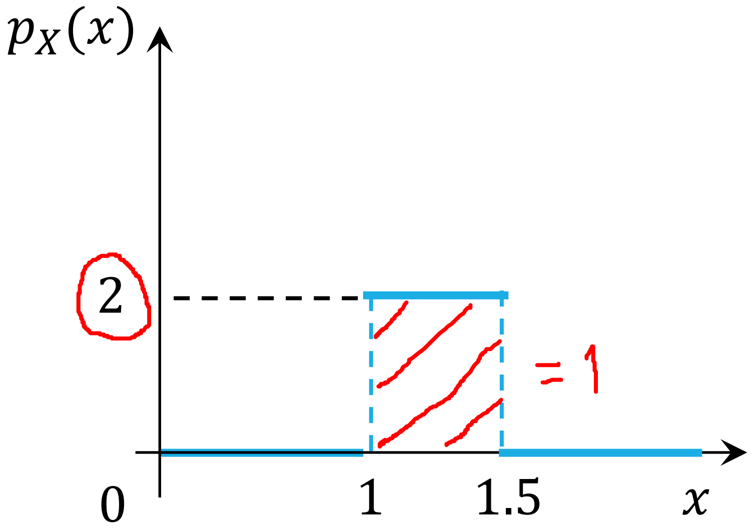
Continuous Uniform Distribution

$$X \sim \text{Uniform}(a,b) \quad p_X(x) = \begin{cases} \dfrac{1}{b-a} & \text{for } x \in [a,b] \\ 0 & \text{otherwise} \end{cases}$$

Takes a random value uniformly in the range $[a,b]$



Example:
$X \sim \text{Uniform}(1, 1.5)$

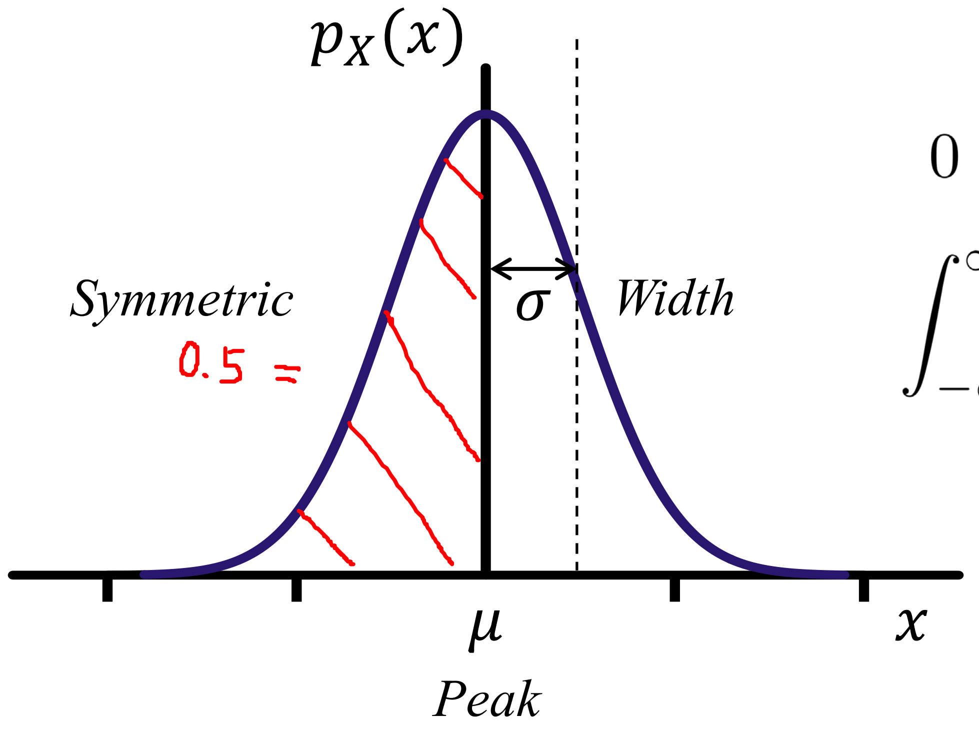


Gaussian (Normal) Distribution

$$X \sim \mathcal{N}(\mu, \sigma^2) \quad p_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$p_X(\mu) = \frac{1}{\sigma\sqrt{2\pi}}$$

Normalization constant



$$0 \leq p_X(x) < \infty$$

$$\int_{-\infty}^{\infty} p_X(x)dx = 1$$

Analogous Continuous Distributions

- Distribution rules still apply to continuous RVs and look similar
- Except **integrals** rather than sums, e.g., for **marginal** PDF:

$$p_X(x) = \int_{-\infty}^{\infty} p_{XY}(x, y)dy$$

"Integrating" out the other variable

Bayes' Rule

- Most important formula in probabilistic machine learning:

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

- Follows directly from product rule:

Set expressions equal and rearrange to derive

$$\Pr(A, B) = \Pr(A|B)\Pr(B)$$
$$\Pr(B, A) = \Pr(B|A)\Pr(A)$$

An Essay towards solving a Problem in the Doctrine of Chances (Thomas Bayes, 1763)

Coding Break

Expectation

- What are μ and σ^2 exactly? $X \sim \mathcal{N}(\mu, \sigma^2)$

E[X] = μ

- Define **expectation** of an RV as:

$$\mathbb{E}[X] = \sum_{x \in \mathcal{X}} x p_X(x) \qquad \mathbb{E}[X] = \int_{-\infty}^{\infty} x p_X(x) dx$$

Discrete *Continuous*

- "Weighted average" of all possible outcomes for an RV

- Properties:

$$\mathbb{E}[aX] = a\,\mathbb{E}[X] \text{ for any } a \in \mathcal{R}$$
$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$$

?: *Linear Operator*

Moments

- Refer to summary statistics as **moments**
- Let the $q \in \mathbb{Z}^+$ moment for a continuous RV be written as:

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

- Can also define **central moments** (shifted about the mean)

$$\mathbb{E}[(X - \mathbb{E}[X])^q] = \int_{-\infty}^{\infty} (x - \mu)^q p_X(x) dx$$

- Recall that variance σ^2 is "spread" or concentration about μ

Variance

- Unique case where $q = 2$ central moment is σ^2:

$$\mathrm{Var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2] = \int_{-\infty}^{\infty} (x-\mu)^2 p_X(x) dx$$

- Alternative expression:

$$\begin{aligned} \mathbb{E}[(X - \mathbb{E}[X])^2] &= \mathbb{E}[X^2 - 2\,\mathbb{E}[X]X + \mathbb{E}[X]^2] \\ &= \mathbb{E}[X^2] - 2\,\mathbb{E}[X]\,\mathbb{E}[X] + \mathbb{E}[X]^2 \\ &= \mathbb{E}[X^2] - \mathbb{E}[X]^2 \end{aligned}$$

$\mathbb{E}[a] = a$ for any $a \in \mathcal{R}$

- Rearrange for 2nd moment:

$$\mathbb{E}[X^2] = \sigma^2 + \mu^2$$

Covariance

- Is a measure of the degree to which two variables are **related**
- Using expectation, the **covariance** for X and Y is defined as:

$$\mathrm{Cov}[X, Y] = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$$

- Can derive: $\mathrm{Cov}[X, Y] = \mathbb{E}[X, Y] - \mathbb{E}[X]\,\mathbb{E}[Y]$

Expectation of joint distribution

$$\mathbb{E}[X, Y] = \mathbb{E}[X]\,\mathbb{E}[Y] \iff X \perp\!\!\!\perp Y$$

Independent

- Properties:

$$\mathrm{Cov}[X, X] = \mathrm{Var}[X]$$

$$X \perp\!\!\!\perp Y \implies \mathrm{Cov}[X, Y] = 0$$

Correlation

- **Pearson correlation coefficient**:

$$\rho = \mathrm{Corr}[X,Y] = \frac{\mathrm{Cov}[X,Y]}{\sigma_X \sigma_Y} \in [-1,1]$$

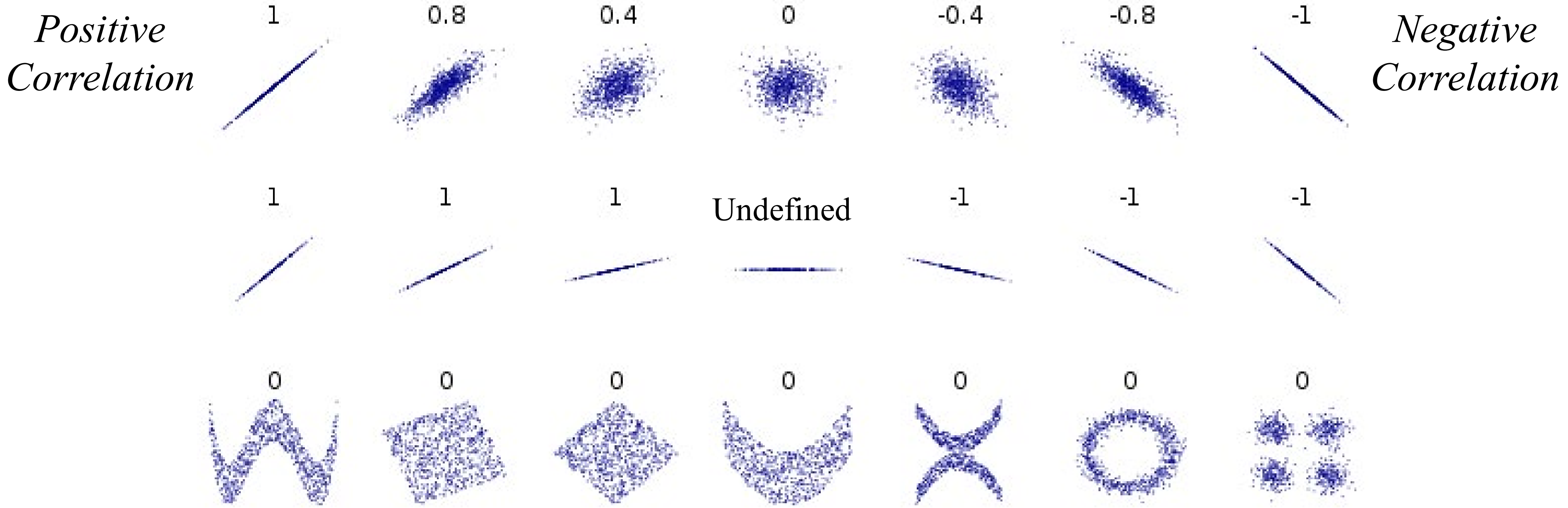
- Normalized measure of covariance
- Independent implies uncorrelated:

$$p_{XY}(X,Y) = p_X(X)p_Y(Y) \implies \mathrm{Corr}[X,Y] = 0$$

- Uncorrelated does NOT imply independent

$$\mathrm{Corr}[X,Y] = 0 \;\not\Longrightarrow\; p_{XY}(X,Y) = p_X(X)p_Y(Y)$$

Visualizing Correlation



Sources: https://en.wikipedia.org/wiki/Pearson_correlation_coefficient

Random Vectors

- Stack n variables into a vector: $\mathbf{x} = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix} \in \mathbb{R}^n$

- Expected value of a random vector:

$$\mathbb{E}[\mathbf{x}] = \int_{-\infty}^{\infty} \mathbf{x}\, p_{X_1,\ldots,X_n}(\mathbf{x}) dx_1 \ldots dx_n$$

$$= \begin{bmatrix} E[X_1] \\ \vdots \\ E[X_n] \end{bmatrix} = \boldsymbol{\mu}$$

Mean Vector

Covariance Matrix

- Is an $n \times n$ square matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{n \times n}$ for $\mathbf{x}$ with entries $\Sigma_{ij} = \mathrm{Cov}[X_i, X_j]$
- Defined as: $\boldsymbol{\Sigma} = \mathrm{Cov}[\mathbf{x}] = \mathbb{E}[(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{x} - \mathbb{E}[\mathbf{x}])^\intercal]$

$$= \begin{bmatrix} \mathrm{Var}[X_1] & \cdots & \mathrm{Cov}[X_1, X_n] \\ \vdots & \ddots & \vdots \\ \mathrm{Cov}[X_n, X_1] & \cdots & \mathrm{Var}[X_n] \end{bmatrix}$$

$$= \mathbb{E}[\mathbf{x}\mathbf{x}^\intercal] - \boldsymbol{\mu}\boldsymbol{\mu}^\intercal$$

$E[xx^T] = \Sigma + \mu\mu^T$

- Useful properties:

$\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^\intercal$ *Symmetric*

$\boldsymbol{\Sigma} \geq 0$ *PSD*

Multivariate Gaussian Distribution

$x \sim N(\mu, \Sigma)$

$$\mathbf{x} \in \mathbb{R}^n \quad \boldsymbol{\mu} \in \mathbb{R}^n \quad \boldsymbol{\Sigma} \in \mathbb{R}^{n \times n}$$

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \quad p_{X_1,\ldots,X_n}(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{n}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\intercal \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

Normalization constant

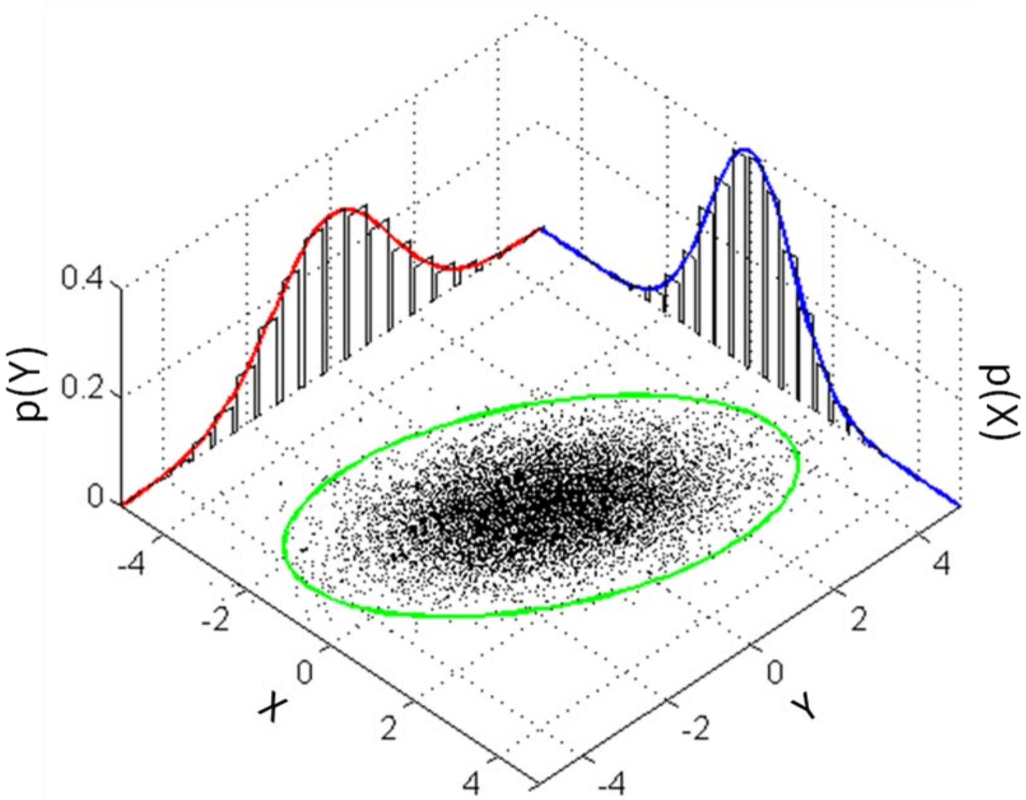
Bivariate Gaussian Distribution

$$\mathbf{x} \in \mathbb{R}^2 \qquad \boldsymbol{\mu} \in \mathbb{R}^2 \qquad \boldsymbol{\Sigma} \in \mathbb{R}^{2\times 2}$$

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}, \boldsymbol{\Sigma} \right)$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_X^2 & \mathrm{Cov}[X,Y] \\ \mathrm{Cov}[Y,X] & \sigma_Y^2 \end{bmatrix} = \begin{bmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{bmatrix}$$

$$\rho = \mathrm{Corr}[X,Y] = \frac{\sigma_{XY}^2}{\sigma_X \sigma_Y}$$



Sources: https://en.wikipedia.org/wiki/Multivariate_normal_distribution

Why Gaussian?

- Only two parameters: μ and σ^2
- **Central Limit Theorem (CLT):** Sum of independent RVs are approximately Gaussian; good choice for modeling "noise"
- Gaussian can be shown to make the "least number of assumptions" (**max entropy**); good default choice
- Analytical form that we can evaluate integrals over
- Lots of nice useful properties…

$$\text{Corr}[X, Y] = 0 \iff X \perp\!\!\!\perp Y$$

Equivalence of uncorrelated and independent

Resources

Probability Review

https://cs229.stanford.edu/section/cs229-prob.pdf

Concluding Remarks

- Look at *"sample_univariate_continuous.ipynb"* notebook on sampling from the univariate uniform and normal continuous distributions:

https://github.com/mazrk7/EECE5644_IntroMLPR_LectureCode/blob/main/notebooks/foundations/sample_univariate_continuous.ipynb

- Also check out *"sample_bivariate_gaussian.ipynb"* for better intuition on a multivariate distribution

https://github.com/mazrk7/EECE5644_IntroMLPR_LectureCode/blob/main/notebooks/foundations/sample_bivariate_gaussian.ipynb

- Questions?