CSE 840: Computational Foundations of Artificial Intelligence March 31, 2025

Bayes' Theorem, Independence, & Discrete Statistical Measures

Instructor: Vishnu Boddeti *Scribe: Molly Thornber*

1 Bayes' Theorem

1.1 Law of Total Probability

Let $B_1, B_2, \ldots, B_k$ be a disjoint partition of Ω with $B_i \in \mathcal{A}$ for all i and $A \in \mathcal{A}$. Then:

$$\begin{aligned} P(A) &= \sum_{i=1}^{k} P(A \mid B_i) \cdot P(B_i) \\ &= \sum_{i=1}^{k} P(A \cap B_i) \end{aligned}$$

Example 1.1.1: Multiple Coin Tosses

Consider tossing a fair coin twice (and let order matter). Let 0 represent getting heads and 1 represent getting tails on a given toss. Intuitively, there are four possible outcomes, represented by the set of ordered tuples $\Omega = \prod_{i=1}^{2} \{0, 1\} = \{(0, 0), (0, 1), (1, 0), (1, 1)\}$.

Let the event $A = \{(0, 1), (1, 1)\}$, i.e. getting tails on the second toss.

We can define $B_1, \ldots, B_k$ as any disjoint partition of Ω, so for intuition's sake, we can use a partition based on the result of the first toss: $B_1 = \{(0, 0), (0, 1)\}$ and $B_2 = \{(1, 0), (1, 1)\}$. Since $B_1 \cup B_2 = \Omega$ and $B_1 \cap B_2 = \emptyset$, this is a disjoint partition.

Finally, since the coin is fair, define P such that each outcome in Ω is equally likely (probability equal to $\frac{1}{4} = 0.25$).

What is the probability of A?

$$\begin{aligned} P(A) &= \sum_{i=1}^{2} P(A \mid B_i) \cdot P(B_i) \\ &= \sum_{i=1}^{k} P(A \cap B_i) \\ &= P(\{(0, 1), (1, 1)\} \cap \{(0, 0), (0, 1)\}) + P(\{(0, 1), (1, 1)\} \cap \{(1, 0), (1, 1)\}) \\ &= P(\{(0, 1)\}) + P(\{(1, 1)\}) \\ &= 0.25 + 0.25 \\ &= 0.5 \end{aligned}$$

Intuitively, we know the probability of getting heads on the second toss is equal to 0.5, so this checks out.

1.2 Bayes' Formula

$$P(B_i \mid A) = \frac{P(A \mid B_i) \cdot P(B_i)}{\sum_{i=1}^{k} P(A \mid B_i) \cdot P(B_i)}$$
$$= \frac{P(A \cap B_i)}{P(A)}$$

Example 1.2.1: COVID Testing

Let COVID status be represented by $C = \{+c, -c\}$ and test result be represented by $T = \{+t, -t\}$.

Assume that:

- 1% of all people have COVID $\quad P(+c) = 0.01$
- 90% of people with COVID test positive (*true positive*) $\quad P(+t \mid +c) = 0.90$
- 8% of people without COVID test positive (*false positive*) $\quad P(+t \mid -c) = 0.08$

Given that a person tested positive, what is the probability that they have COVID?

$$P(+c \mid +t) = \frac{P(+t \mid +c) \cdot P(+c)}{P(+t \mid +c) \cdot P(+c) + P(+t \mid -c) \cdot P(-c)}$$
$$= \frac{0.9 \cdot 0.01}{0.9 \cdot 0.01 + 0.08 \cdot 0.99}$$
$$\approx 10\%$$

2 Independence

2.1 Independence of Events & Families of Events

Definition 1 *Consider a probability space* $(\Omega, \mathcal{A}, P)$.

*Two events A and B are called **independent** ($A \perp\!\!\!\perp B$) if:*

$$P(A \cap B) = P(A) \cdot P(B)$$

*A family of events $(A_i)_{i \in I}$ is called (mutually) **independent** if for all finite subsets $J \subseteq I$ we have:*

$$P\left(\bigcap_{i \in J} A_i\right) = \prod_{i \in J} P(A_i)$$

*A family of events $(A_i)_{i \in I}$ is called **pairwise independent** if $\forall\, i, j \in I$:*

$$P(A_i \cap A_j) = P(A_i) \cdot P(A_j)$$

Note that pairwise independence **does not imply** independence, but independence **does imply** pairwise independence:

$$\text{pairwise independence} \quad \begin{matrix} \not\Rightarrow \\ \Leftarrow \end{matrix} \quad \text{independence}$$

Observation 2 *For two events A and B,*

$$A \perp\!\!\!\perp B \;\; \Leftrightarrow \;\; P\left(A \mid B\right) = P\left(A\right)$$

Example 2.1.1: Coin Tosses & Independence

Consider the same probability space as Example 1.1.1.

First, let events A and B represent getting tails on the first and second flips, respectively, of the fair coin. Based on the probability space, we know that $P\left(A\right) = P\left(B\right) = \frac{2}{4} = 0.5$ and that $P\left(A \cap B\right) = \frac{1}{4} = 0.25$.

We can show that A and B are **independent**:

$$P\left(A\right) \cdot P\left(B\right) = 0.5 \cdot 0.5 = 0.25 = P\left(A \cap B\right)$$

Now, consider a third variable C, representing the event in which exactly one of the two coin tosses was tails (i.e. $C = A \oplus B$). Now, we have a new event space:

$$\Omega = \left\{\left(0,\, 0,\, 0\right),\, \left(0,\, 1,\, 1\right),\, \left(1,\, 0,\, 1\right),\, \left(1,\, 1,\, 0\right)\right\}$$

We can show that the family of events $X = \{A,\, B,\, C\}$ is **pairwise independent**:

$$\begin{aligned} P\left(A\right) \cdot P\left(B\right) &= 0.5 \cdot 0.5 = 0.25 & &= P\left(A \cap B\right) = 0.25 \\ P\left(A\right) \cdot P\left(C\right) &= 0.5 \cdot 0.5 = 0.25 & &= P\left(A \cap C\right) = 0.25 \\ P\left(B\right) \cdot P\left(C\right) &= 0.5 \cdot 0.5 = 0.25 & &= P\left(B \cap C\right) = 0.25 \end{aligned}$$

We can also show that this family, X, is **not independent**:

$$\begin{aligned} P\left(A\right) \cdot P\left(B\right) \cdot P\left(C\right) = 0.5 \cdot 0.5 \cdot 0.5 &= 0.125 \\ P\left(A \cap B \cap C\right) &= 0 \end{aligned}$$

$$P\left(A\right) \cdot P\left(B\right) \cdot P\left(C\right) \neq P\left(A \cap B \cap C\right)$$

2.2 Independence of Random Variables

Definition 3 *Two* <u>*random variables*</u> *$X : \Omega \to \Omega_1$ and $Y : \Omega \to \Omega_2$ are called **independent** ($X \perp\!\!\!\perp Y$) if their induced σ-algebras $\sigma\left(X\right)$ and $\sigma\left(Y\right)$ are independent:*

$$\forall\, A \in \sigma\left(X\right),\, B \in \sigma\left(Y\right) \;:\; P\left(A \cap B\right) = P\left(A\right) \cdot P\left(B\right)$$

2.3 Independence: Key Concepts

- **Probability**:
  - Central Limit Theorem (CLT): for independent random variables $\{X_1, \ldots, X_n\}$, the sample means will converge to the expected population mean as $n \to \infty$
  - Addition of random variables: addition of independent random variables holds certain properties such as
    - $E(X + Y) = E(X) + E(Y)$
    - $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$
- **Algorithmic fairness/invariance**: fairness can be achieved by enforcing $\widehat{y} \perp\!\!\!\perp S$ while learning for some predicted label $\widehat{y}$ and demographic variable S
- **Learning theory**: proving that an algorithm learns the correct prediction, converges at a certain rate, etc. often requires the assumption that the training samples are independent

3 Expectation (Discrete Case)

3.1 Expectation

Definition 4 *Let $(\Omega, \mathcal{A}, P)$ be a probability space, $S \subset \mathbb{R}$ be at most countable, and $X : \Omega \to S$ be a discrete random variable (i.e. image $X(\Omega)$ is at most countable).*

If $\sum_{r \in S} |r| \cdot P(X = r) < \infty$, then

$$E(X) := \sum_{r \in S} r \cdot P(X = r)$$

*is called the **expectation** of X.*

Note: May also be written as $E\,X$, $\mathbb{E}\,X$, or $\mathbb{E}(X)$

Note: Equivalent to the weighted mean of the possible values of X, or the expected mean of X

Example 3.1.1: Coin Toss & Expectation

Consider tossing one coin. The sample space is $\Omega = \{\text{HEADS}, \text{TAILS}\}$ and the event space is $\mathcal{A} = \mathcal{P}(\Omega)$ (a power series). We can define a variable $0 < p < 1$ such that $P(\text{HEADS}) = p$ and $P(\text{TAILS}) = 1 - p$. Finally, we can define $X : \Omega \to \{0, 1\}$ such that $\text{HEADS} \mapsto 0$, $\text{TAILS} \mapsto 1$. Then

$$\begin{aligned} E(X) &= 0 \cdot P(X = 0) + 1 \cdot P(X = 1) \\ &= 0 \cdot P(\text{HEADS}) + 1 \cdot P(\text{TAILS}) \\ &= 0 \cdot p + 1 \cdot (1 - p) \\ &= 1 - p \end{aligned}$$

Example 3.1.2: Classifier Error

Let $\widehat{y} = f(x)$ where f is a classifier, x is the input, and $\widehat{y}$ is the classifier output. Let y be the target output. Then the classification error can be calculated using

$$e = (\widehat{y} - y)^2 = (f(x) - y)^2$$

We can minimize error using

$$\min_f E_X(e)$$

Example 3.1.3: Lists of Numbers (Expectation)

Let L_1 and L_2 be list of 10 numbers each:

$$\begin{aligned} L_1 &= [1,\ 2,\ 2,\ 3,\ 3,\ 3,\ 4,\ 4,\ 4,\ 4] \\ L_2 &= [1,\ 1,\ 2,\ 2,\ 3,\ 3,\ 4,\ 4,\ 5,\ 5] \end{aligned}$$

Let X be a random variable representing the value of an element chosen randomly from L_1. Then

$$\begin{aligned} E(X) &= 1 \cdot P(X = 1) + 2 \cdot P(X = 2) + 3 \cdot P(X = 3) + 4 \cdot P(X = 4) \\ &= 1 \cdot \frac{1}{10} + 2 \cdot \frac{2}{10} + 3 \cdot \frac{3}{10} + 4 \cdot \frac{4}{10} \\ &= 0.1 + 0.4 + 0.9 + 1.6 \\ &= 3 \end{aligned}$$

Let Y be a random variable representing the value of an element chosen randomly from L_2. Then

$$\begin{aligned} E(Y) &= 1 \cdot P(Y = 1) + 2 \cdot P(Y = 2) + 3 \cdot P(Y = 3) + 4 \cdot P(Y = 4) + 5 \cdot P(Y = 5) \\ &= 1 \cdot \frac{2}{10} + 2 \cdot \frac{2}{10} + 3 \cdot \frac{2}{10} + 4 \cdot \frac{2}{10} + 5 \cdot \frac{2}{10} \\ &= 0.2 + 0.4 + 0.6 + 0.8 + 1.0 \\ &= 3 \end{aligned}$$

3.2 Centered Random Variables

Definition 5 *A random variable X is called **centered** if*

$$E(X) = 0$$

3.3 Properties of Expectation

Let X and Y be random variables.

- **Independence & expectation**:

$$X \perp\!\!\!\perp Y \;\;\Rightarrow\;\; E(X \cdot Y) = E(X) \cdot E(Y)$$

- **Linearity**: (for $a,\, b \in \mathbb{R}$)

$$\begin{aligned} E(a \cdot X + b \cdot Y) &= a \cdot E(X) + b \cdot E(Y) \\ E(a \cdot X + b) &= a \cdot E(X) + b \\ E(a) &= a \end{aligned}$$

4 Variance, Covariance, & Correlation (Discrete Case)

Let $X,\, Y : (\Omega,\, \mathcal{A},\, P) \to \mathbb{R}$ be discrete random variables with $E\left(X^2\right) < \infty$ and $E\left(Y^2\right) < \infty$.

4.1 Variance

Definition 6 *The* ***variance*** *of X is defined as*

$$\mathrm{Var}(X) := E\left((X - E(X))^2\right)$$

Definition 7 *The* ***standard deviation*** *of X is defined as*

$$\sigma_X := \sqrt{\mathrm{Var}(X)}$$

Example 4.1.1: Lists of Numbers (Variance)

Consider random variables X and Y as defined in 3.1.3.

The **variance** and **standard deviation** of X are

$$\begin{aligned} \mathrm{Var}(X) &= E\left((X - E(X))^2\right) \\ &= E\left(X^2 - 2 \cdot X \cdot E(X) + E(X)^2\right) \\ &= E\left(X^2\right) - 2 \cdot E(X)^2 + E(X)^2 \\ &= E\left(X^2\right) - E(X)^2 \\ &= \left(1^2 \cdot \frac{1}{10} + 2^2 \cdot \frac{2}{10} + 3^2 \cdot \frac{3}{10} + 4^2 \cdot \frac{4}{10}\right) - 3^2 \\ &= 0.1 + 0.8 + 2.7 + 6.4 - 9 \\ &= 1 \end{aligned}$$

$$\begin{aligned} \sigma_X &= \sqrt{\mathrm{Var}(X)} \\ &= \sqrt{1} = 1 \end{aligned}$$

The **variance** and **standard deviation** of Y are

$$\begin{aligned}\mathrm{Var}\,(Y) &= E\left((Y - E\,(Y))^2\right) \\ &\dots \\ &= E\left(Y^2\right) - E\,(Y)^2 \\ &= \left(1^2 \cdot \frac{2}{10} + 2^2 \cdot \frac{2}{10} + 3^2 \cdot \frac{2}{10} + 4^2 \cdot \frac{2}{10} + 5^2 \cdot \frac{2}{10}\right) - 3^2 \\ &= 0.2 + 0.8 + 1.8 + 3.2 + 5 - 9 \\ &= 2\end{aligned}$$

$$\begin{aligned}\sigma_Y &= \sqrt{\mathrm{Var}\,(Y)} \\ &= \sqrt{2}\end{aligned}$$

4.2 Covariance

Definition 8 *The* ***covariance*** *of X and Y is defined as*

$$\mathrm{Cov}\,(X,\,Y) := E\,((X - E\,(X)) \cdot (Y - E\,(Y)))$$

Example 4.2.1: Lists of Numbers (Covariance)

Consider random variables X and Y as defined in 3.1.3.

The **covariance** of X and Y is

$$\begin{aligned}\mathrm{Cov}\,(X,\,Y) &= E\,((X - E\,(X)) \cdot (Y - E\,(Y))) \\ &= E\,(X \cdot Y - X \cdot E\,(Y) - Y \cdot E\,(X) + E\,(X) \cdot E\,(Y)) \\ &= E\,(X \cdot Y) - 2 \cdot E\,(X) \cdot E\,(Y) + E\,(X) \cdot E\,(Y) \\ &= E\,(X \cdot Y) - E\,(X) \cdot E\,(Y) \\ &= \left(1 \cdot \frac{1}{10} + 2 \cdot \frac{1}{10} + 4 \cdot \frac{1}{10} + 6 \cdot \frac{1}{10} + 9 \cdot \frac{2}{10} + 16 \cdot \frac{2}{10} + 20 \cdot \frac{2}{10}\right) - 3 \cdot 3 \\ &= 0.1 + 0.2 + 0.4 + 0.6 + 1.8 + 3.2 + 4.0 - 9 \\ &= 1.3\end{aligned}$$

4.3 Correlation

Definition 9 *The* ***correlation coefficient*** *between X and Y is defined as*

$$\rho_{XY} = \frac{\mathrm{Cov}\,(X,\,Y)}{\sigma_X \cdot \sigma_Y} \in [-1,\,1]$$

Definition 10 *If* $\text{Cov}(X, Y) = 0$*, then* X *and* Y *are called* ***uncorrelated****.*

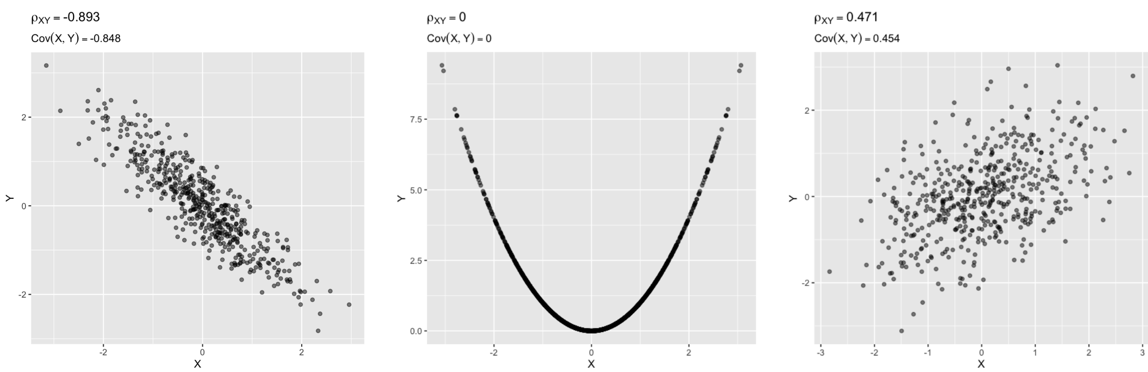
Example 4.3.1: Lists of Numbers (Correlation)

Consider random variables X and Y as defined in 3.1.3.

The **correlation coefficient** between X and Y is

$$\begin{aligned}\rho_{XY} &= \frac{\text{Cov}(X, Y)}{\sigma_X \cdot \sigma_Y} \\ &= \frac{1.3}{1 \cdot \sqrt{2}} \\ &\approx 0.91924\end{aligned}$$

4.3.1 Intuition About Correlation & Covariance



(a) $\text{Cov}(X, Y) < 0$ and $\rho_{XY} < 0$ (b) $\text{Cov}(X, Y) = 0$ and $\rho_{XY} = 0$ (c) $\text{Cov}(X, Y) > 0$ and $\rho_{XY} > 0$

Figure 1: Varying signs and magnitudes of covariance and correlation coefficient

- The sign (positive, negative, or zero) of the covariance will be the same as the sign of the correlation coefficient:

$$\begin{aligned}\text{Cov}(X, Y) < 0 \quad &\Rightarrow \quad \rho_{XY} < 0 \\ \text{Cov}(X, Y) = 0 \quad &\Rightarrow \quad \rho_{XY} = 0 \\ \text{Cov}(X, Y) > 0 \quad &\Rightarrow \quad \rho_{XY} > 0\end{aligned}$$

- In general (but not always, since ρ also depends on σ), a higher magnitude (absolute value) covariance is associated with a higher magnitude ρ

- Independence & correlation:

$$\rho_{XY} = 0 \quad \overset{\not\Rightarrow}{\Leftarrow} \quad X \perp\!\!\!\perp Y$$

- Independence & covariance:

$$\text{Cov}(X, Y) = 0 \quad \overset{\not\Rightarrow}{\Leftarrow} \quad X \perp\!\!\!\perp Y$$

4.4 Moments

For $k \in \mathbb{N}$:

Definition 11 *The* **k-*th moment*** *of* X *is defined as*

$$E\left(X^k\right)$$

Using this definition:

- $k = 0: \quad E\left(X^0\right) = 1$
- $k = 1: \quad E\left(X^1\right) = E\left(X\right)$ (the expectation of X)

Definition 12 *The* **k-*th centered moment*** *of* X *is defined as*

$$E\left(\left(X - E\left(X\right)\right)^k\right)$$

Using this definition:

- $k = 0: \quad E\left(\left(X - E\left(X\right)\right)^0\right) = 1$
- $k = 1: \quad E\left(\left(X - E\left(X\right)\right)^1\right) = E\left(X\right) - E\left(X\right) = 0$
- $k = 2: \quad E\left(\left(X - E\left(X\right)\right)^2\right) = \mathrm{Var}\left(X\right)$ (the variance of X)

4.5 Properties

- $\mathrm{Var}\left(X\right) = E\left(X^2\right) - \left(E\left(X\right)\right)^2$
- $\mathrm{Cov}\left(X, Y\right) = E\left(X \cdot Y\right) - E\left(X\right) \cdot E\left(Y\right)$
- $E\left(a \cdot X + b\right) = a \cdot E\left(X\right) + b$
- $\mathrm{Var}\left(a \cdot X + b\right) = a^2 \cdot \mathrm{Var}\left(X\right)$
- $\mathrm{Cov}\left(X, Y\right) = \mathrm{Cov}\left(Y, X\right)$
- $\mathrm{Var}\left(X + Y\right) = \mathrm{Var}\left(X\right) + \mathrm{Var}\left(Y\right) + 2 \cdot \mathrm{Cov}\left(X, Y\right)$
- $X \perp\!\!\!\perp Y \quad \Rightarrow \quad \mathrm{Var}\left(X + Y\right) = \mathrm{Var}\left(X\right) + \mathrm{Var}\left(Y\right)$