COE548: LARGE LANGUAGE MODELS

Topic: Probability Theory & Statistics

الجامعة اللبنانيّة الأميركيّة
Lebanese American University

Outline

Probability Theory & Statistics

- Conditional Probability and Independence
- Random Variables
- Cumulative Distribution Function (CDF), Probability Mass Function (PMF), Probability Density Function (PDF)
- Expectation and Variance
- Multivariate Probability
- Baye’s Theorem
- Statistics
- Maximum Likelihood Estimation (MLE)
- Markov Chains/Hidden Markov Models (HMMs)

Probability Theory

- Basic elements of probability theory:
  - *Sample space:* Ω *(The set of all random outcomes that can happen)*
  - *Event:* $A \subseteq \Omega$ *(Subset of the sample space)*
  - *Event space:* F *(The power set of events)*
  - *Probability measure:* $P: F \to \mathbb{R}$ *(Takes as input an event and assigns it a value between 0 and 1)*
- Three axioms of probability:
  - $P(A) \geq 0, \ \forall A \in F$
  - $P(\Omega) = 1$
  - *If* $A_1, A_2, \ldots$ *disjoint sets of events then* $P \bigcup_i A_i = \sum_i A_i$

Conditional Probability and Independence

- Let B be any event such that $P(B) \neq 0$.
- It stands that $P(A|B) = \frac{P(A \cap B)}{P(B)}$.
- A and B are independent ($A \perp B$) if and only if $\mathrm{P}(A \cap \mathrm{B}) = \mathrm{P(A)P(B)}$, which means $P(A|B) = P(A)$ (probability of A does not change whether B occurred or not).

Random Variables

- A random variable $X: \Omega \rightarrow \mathbb{R}$ (Mapping from outcomes to real-values)
  - *Example (coin toss):* $w_o = HHHTHTTHTT$, *then:*
    - the number of heads is $X(w_o) = 5$.
    - Number of tosses until tails is $X(w_o) = 4$.
  - $Val(X) = X(\Omega)$ *(value of* X*)*
- We assign probabilities to events, not outcomes.
- By defining random variables we bring random unordered events to the real line, which makes it mathematically manipulatable.
- The probability distribution describes how the probabilities are distributed over the values of the random variable. For a discrete random variable, this is often represented by the probability mass function (PMF), which gives the probability of each specific outcome.

Example: Coin Toss

1. **Sample Space**: Imagine all possible outcomes of an experiment. For the die, the sample space is {1,2,3,4,5,6}. These are all the events that could happen.
2. **Random Variable as a Mapping**: The random variable X takes each event in the sample space and maps it to a real number. In this simple case, it's a straightforward mapping: $X(1) = 1$, $X(2) = 2$ and so on. But in more complex scenarios, the mapping might not be as direct.
3. **Probability of an Event**: The probability of an event (say, rolling a 4) is the likelihood of the event occurring. Since $X(4) = 4$, the probability that $X = 4$ is 1/6.
4. **Probability Distribution**: If you think of all possible values that X can take and the corresponding probabilities, you have a probability distribution. For the die roll, the probability distribution tells you that $X = 1$ with probability $1/6$, $X = 2$ with probability $1/6$, and so on.

Discrete vs Continuous RVs

Discrete RV

- $Val(X)$ is countable

$$P(X = k) = P(\{\omega | X(\omega) = k\})$$

- Probability Mass Function (PMF)

$$\rho_X : Val(X) \to [0,1]$$
$$\rho_X = P(X = x)$$
$$\sum_{x \in Val(X)} \rho_X = 1$$

Continuous RV

- $Val(X)$ is uncountable

$$P(a \leq X \leq b) = P(\{\omega | a \leq X(\omega) \leq b\})$$

- Probability Density Function (PDF)

$$f_X : \mathbb{R} \to \mathbb{R}$$
$$f_X(X) = \frac{d}{dx} F_X(x)$$
$$f_X(X) \neq P(X = x)$$
$$\int_{-\infty}^{\infty} f_X(x) dx$$

Cumulative Distribution Function (CDF)

- The cumulative distribution function (CDF) of a random variable X is a function that tells us the probability that X takes a value less than or equal to a certain value. It accumulates (or sums up) the probabilities up to that point.

$$F_X\colon \mathbb{R} \rightarrow [0,1]$$
$$F_X(x) = P(X \leq x) = P(\{\omega | X(\omega) \leq x\})$$

- Example: If X is the outcome of a die roll, the CDF $\mathrm{F}_X(3)$ would give the probability that you roll a 1, 2, or 3. In other words:

$$F_X(3) = P(X \leq 3) = P(X = 1) + P(X = 2) + P(X = 3) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = 0.5$$

Note: The PDF is the derivative of the CDF.

Expected Value

- Expected value represents the "average" or "mean" value of the possible outcomes of a random variable, weighted by their probabilities. It provides a measure of the central tendency of a probability distribution.
- Let $g: \mathbb{R} \to \mathbb{R}$, X be a discrete RV with PMF p_X, then the expectation of $g(x)$ is

$$\mathbb{E}[g(x)] = \sum_{x \in Val(X)} g(x)\rho_X(x)$$

For continuous: $\mathbb{E}[g(x)] = \int_{-\infty}^{\infty} g(x) f_X(x)$.

- Back to rolling dice example:

$$\mathbb{E}[X] = 1\frac{1}{6} + 2\frac{1}{6} + 3\frac{1}{6} + 4\frac{1}{6} + 5\frac{1}{6} + 6\frac{1}{6} = 3.5$$

Law of Large Numbers

Assuming

$$E[X] = \lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} g\left(x^{(i)}\right) = \int_{-\infty}^{\infty} g(x) f_X(x) dx$$

- As N increases the summation approaches the true expectation almost surely (with probability 1).
- Without taking the limit, we call $\frac{1}{N}\sum_{i=1}^{N} g\left(x^{(i)}\right)$ the Monte Carlo estimate.

Variance

- Variance is a statistical measure that tells us how much the values of a random variable differ from the expected value of that variable.

$$Var(X) = \mathbb{E}[(X - \mathbb{E}[X])^2]$$

Or an alternative formula:

$$Var(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

- Back to rolling dice example given expected value is 3.5:

$$Var(X) = \sum_{i=1}^{6} (x_i - 3.5)^2 \frac{1}{6}$$

$$= \frac{(1-3.5)^2 + (2-3.5)^2 + (3-3.5)^2 + (4-3.5)^2 + (5-3.5)^2 + (6-3.5)^2}{6} \approx 2.92$$

Parameters

Parameters

Distribution	PDF or PMF	Mean	Variance
Bernoulli(p)	$\begin{cases} p, & \text{if } x = 1 \\ 1-p, & \text{if } x = 0. \end{cases}$	p	$p(1-p)$
Binomial(n, p)	$\binom{n}{k} p^k (1-p)^{n-k}$ for $k = 0, 1, \ldots, n$	np	$np(1-p)$
Geometric(p)	$p(1-p)^{k-1}$ for $k = 1, 2, \ldots$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Poisson(λ)	$\frac{e^{-\lambda}\lambda^k}{k!}$ for $k = 0, 1, \ldots$	λ	λ
Uniform(a, b)	$\frac{1}{b-a}$ for all $x \in (a, b)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Gaussian(μ, σ^2)	$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ for all $x \in (-\infty, \infty)$	μ	σ^2
Exponential(λ)	$\lambda e^{-\lambda x}$ for all $x \geq 0, \lambda \geq 0$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$

Two Random Variables

- Bivariate CDF: $F_{XY}(x, y) = P(X \leq x, Y \leq y)$

- Bivariate PMF: $\rho_{XY}(x, y) = P(X = x, Y = y)$
- Marginal PMF: $\rho_X(x) = \sum_y \rho_{XY}(x, y)$

- Bivariate PDF: $f_{XY}(x, y) = \frac{\partial^2 F_{XY}(x,y)}{\partial x \partial y}$
- Marginal PDF: $f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x, y) dy$

Two Random Variables

Example with two die: Let X be the outcome of the first die and Y be the outcome of the second die.

- **Joint PMF**: The joint PMF $P(X = x, Y = y)$ for each possible pair (x, y) is $\frac{1}{36}$, because there are 36 equally likely outcomes when rolling two dice.
- **Marginal PMF**: The marginal PMF for X would be:

$$P(X = x) = \sum_{y=1}^{6} P(X = x, Y = y) = \sum_{y=1}^{6} \frac{1}{36} = \frac{1}{6}$$

- **Joint CDF**: $F_{XY}(X \leq 3, Y \leq 4)$ is equal to the sum of probabilities of outcomes where $X \leq 3, Y \leq 4$. There are 12 such pairs, so: $F_{XY}(3{,}4) = \frac{12}{36} = \frac{1}{3}$

Independence of Two RVs

- Two random variables X and Y are independent if:

$$\rho_{XY}(x, y) = \rho_X(x)\rho_Y(y)$$

And

$$\rho_{Y|X}(x, y) = \rho_Y(y)$$

Bayes' Theorem

- Given the conditional probability of an event, $P(x|y)$, we want to find the reverse conditional probability:

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)}$$

Where $P(x) = \sum_{y' \in Val(Y)} P(x|y')P(y')$.

- Note, the joint is the product of the marginal and conditional (this is by the chain rule of probability theory):

$$\underbrace{P(x,y)}_{\text{Joint}} = \underbrace{P(x)}_{\text{Marginal}}\underbrace{P(y|x)}_{\text{Conditional}} = P(y)P(x|y)$$

Bayes' Theorem

By rearranging the terms we get:

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)}$$

Commonly written as:

$$P(y|x) = \frac{P(y)P(x|y)}{\sum_{y' \in Val(Y)} P(y')P(x|y')} = \frac{P(y)P(x|y)}{\sum_{y' \in Val(Y)} P(x, y')}$$

Expectation of Two RVs

- Discrete case:

$$\mathbb{E}[XY] = \sum_i \sum_j x_i, y_j P\big(X = x_i, Y = y_j\big)$$

$$\mathbb{E}[g(x,y)] = \sum_i \sum_j g(x_i, y_j) P\big(X = x_i, Y = y_j\big)$$

- Continuous case:

$$\mathbb{E}[XY] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy f_{XY}(x,y)\, dx\, dy$$

$$\mathbb{E}[g(x,y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y) f_{XY}(x,y)\, dx\, dy$$

Covariance of Two RVs

- Covariance is a measure of how much two random variables change together.
- If the variables tend to increase together, the covariance is positive.
- If one tends to increase when the other decreases, the covariance is negative.
- If the variables are independent, the covariance is zero (though a covariance of zero does not necessarily imply independence).

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

- If X and Y are independent then:

$$\mathbb{E}[XY] = E[X]E[Y] \rightarrow Cov[x, y] = 0$$

$$Var[X + Y] = \mathbb{E}[(X + Y)^2] - \mathbb{E}[(X + Y)]^2$$

$$Var[X + Y] = Var[X] + Var[Y] + 2Cov[X, Y]$$

Multivariate Gaussian Distribution

- Say we are dealing with vectors $x \in \mathbb{R}^n$ and we want to model $P(x_1), P(x_2), \ldots$ with the parameters:
  - *Expected value:* $\mu \in \mathbb{R}^n$
  - *Covariance:* $\Sigma \in \mathbb{R}^{n \times n}$

$$P(x; \mu, \Sigma) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right)$$

Note:

Σ: is a positive semi-definite matrix, that is full-rank (inverse exists)

$(x-\mu)^T \Sigma^{-1} (x-\mu)$: is quadratic form

Conditional Probability and Expectation

- Let X, Y be RVs with the same probability space.

$$P(X = x|Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)}$$

The conditional expectation given by:

$$\mathbb{E}[X|Y = y] = \sum_x x \frac{P(X = x, Y = y)}{P(Y = y)}$$

Is a random variable itself. Explanation:

- The conditional expectation $\mathbb{E}[X|Y = y]$ is the expected value of X given that the random variable Y is equal to some specific value y. This is a number (not a random variable) when y is fixed.
- However, when we look at the conditional expectation as a function of Y (without fixing y), it is itself a random variable. Specifically, the conditional expectation $\mathbb{E}[X|Y]$ is a function that depends on Y and varies as Y varies.

Conditioned Bayes' Rule

$$P(a|b,c) = \frac{P(b|a,c)P(a|c)}{P(b|c)}$$

Proof:

$$\frac{P(b|a,c)P(a|c)}{P(b|c)} = \frac{P(b,a,c)P(a|c)}{P(b|c)P(a,c)}$$

$$= \frac{P(b,a,c)P(a,c)}{P(b|c)P(a,c)P(c)} = \frac{P(b,a,c)}{P(b|c)P(c)}$$

$$= \frac{P(b,a,c)}{P(b,c)} = P(a|b,c)$$

Statistics vs ML

- Given some data (observations) $x \in \mathbb{R}^n$ you are trying to infer the parameters that define the probability distribution over that data.
- Many techniques to estimate parameters given data:
  - *Method of moments*
  - *Maximum likelihood estimation*
  - *Etc.*
- Relation to ML:
  - *Using (training) data, we estimate parameters of a model to make predictions on future data.*
  - *In statistics the goal is to make statements about the probability distribution parameters, whereas in ML we only care about model accuracy/performance.*

Maximum Likelihood Estimation

- Assumption: Gaussian data
- Have a collection of i.i.d data $x \in \mathbb{R}^d$ with n samples, i.e., $x^{(1)}, \dots, x^{(n)} \in \mathbb{R}^d$.

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

Recall if $x^{(1)}, \dots, x^{(n)}$ are independent then

$$P\left(x^{(1)}, \dots, x^{(n)}\right) = \mathrm{P}\left(x^{(1)}\right) P\left(x^{(2)}\right) \dots P\left(x^{(n)}\right) = \prod_{i=1}^{n} P\left(x^{(i)}\right)$$

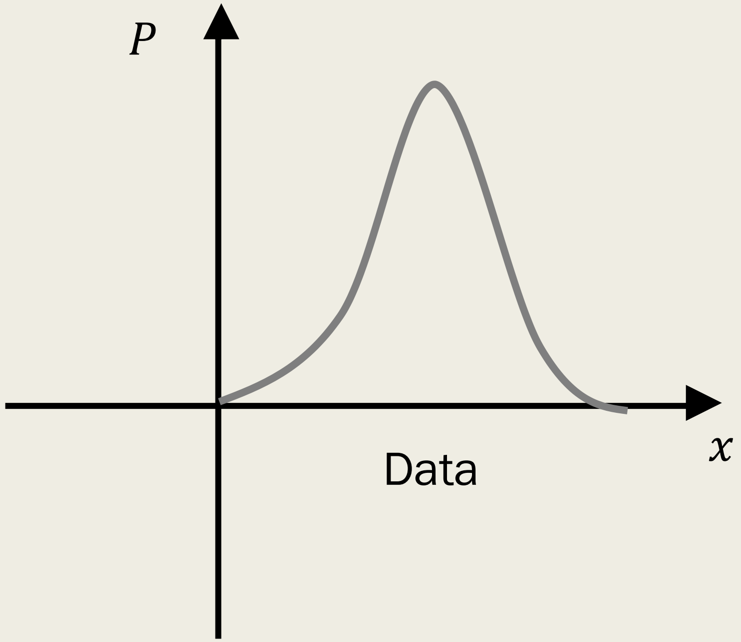
Then we write:

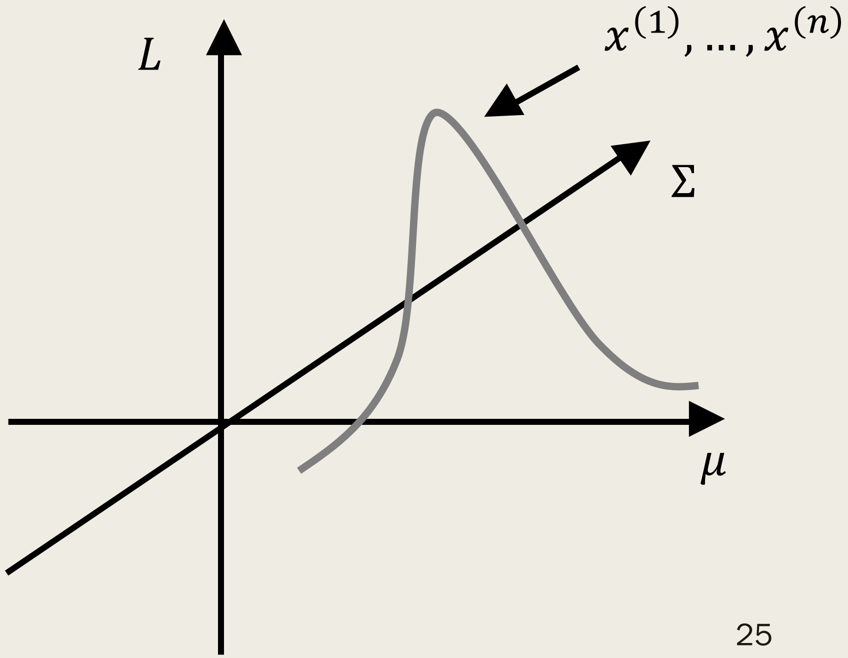
$$P(x; \mu, \Sigma) = \prod_{i=1}^{n} \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\left(x^{(i)}-\mu\right)^T \Sigma^{-1} \left(x^{(i)}-\mu\right)\right)$$

Maximum Likelihood Estimation

Define a likelihood function (of the parameters):

$$L(\mu, \Sigma; x^{(1)}, \dots, x^{(n)}) = \prod_{i=1}^{n} \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\left(x^{(i)} - \mu\right)^T \Sigma^{-1} \left(x^{(i)} - \mu\right)\right)$$





Maximum Likelihood Estimation

$$L(\theta; x) = \prod_{i=1}^{n} L\left(\theta; x^{(i)}\right)$$

And in MLE we want:

$$\hat{\theta}_{MLE} = \operatorname*{argmax}_{\theta} \prod_{i=1}^{n} L\left(\theta; x^{(i)}\right)$$

Equivalently:

$$\hat{\theta}_{MLE} = \operatorname*{argmax}_{\theta} \log \prod_{i=1}^{n} L\left(\theta; x^{(i)}\right) = \operatorname*{argmax}_{\theta} \sum_{i=1}^{n} \log L\left(\theta; x^{(i)}\right)$$

Recall: log is a monotonically increasing function.

Maximum Likelihood Estimation

$$\hat{\theta}_{MLE} = \underset{\theta}{\operatorname{argmax}} \log \prod_{i=1}^{n} L\left(\theta; x^{(i)}\right) = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{n} \log L\left(\theta; x^{(i)}\right)$$

If likelihood is Gaussian:

$$\hat{\theta}_{MLE} = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{n} \log \left[\frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} \left(x^{(i)} - \mu\right)^T \Sigma^{-1} \left(x^{(i)} - \mu\right) \right) \right]$$

Thus we have

$$\hat{\theta}_{MLE} = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{n} K - \frac{1}{2} \log |\Sigma| - \frac{1}{2} \left(x^{(i)} - \mu\right)^T \Sigma^{-1} \left(x^{(i)} - \mu\right)$$

To get the argmax we take the derivative and set it to zero.

Maximum Likelihood Estimation

Recall: $\nabla_x x^T A x = 2Ax$

$$\hat{\mu}_{MLE} = \nabla_\mu \sum_{i=1}^{n} K - \frac{1}{2}\log|\Sigma| - \frac{1}{2}\left(x^{(i)} - \mu\right)^T \Sigma^{-1}\left(x^{(i)} - \mu\right)$$

$$= \nabla_\mu - \frac{1}{2}\sum_{i=1}^{n}\left[x^{(i)^T}\Sigma^{-1}x^{(i)} - x^{(i)^T}\Sigma^{-1}\mu - \mu^T\Sigma^{-1}x^{(i)} + \mu^T\Sigma^{-1}\mu\right]$$

$$= \nabla_\mu - \frac{1}{2}\sum_{i=1}^{n}\left[-2\left(\Sigma^{-1}x^{(i)}\right)^T\mu + \mu^T\Sigma^{-1}\mu\right]$$

$$= \sum_{i=1}^{n}\left[\Sigma^{-1}x^{(i)} - \Sigma^{-1}\mu\right] = 0$$

Therefore, $\hat{\mu}_{MLE} = \frac{1}{n}\sum_{i=1}^{n} x^{(i)}$

Maximum Likelihood Estimation

Recall: $\nabla_x x^T A x = 2Ax$

Note: For $S = \Sigma^{-1}$ *then*$\nabla_S l = 0 \Leftrightarrow \nabla_{\Sigma}^{-1} l = 0$*, and* $\nabla_A x^T A x = x x^T$

$$\nabla_\Sigma \sum_{i=1}^{n} K - \frac{1}{2}\log|\Sigma| - \frac{1}{2}\left(x^{(i)} - \mu\right)^T \Sigma^{-1}\left(x^{(i)} - \mu\right)$$

Becomes

$$\nabla_S \sum_{i=1}^{n} K - \frac{1}{2}\log|S^{-1}| - \frac{1}{2}\left(x^{(i)} - \mu\right)^T S\left(x^{(i)} - \mu\right)$$

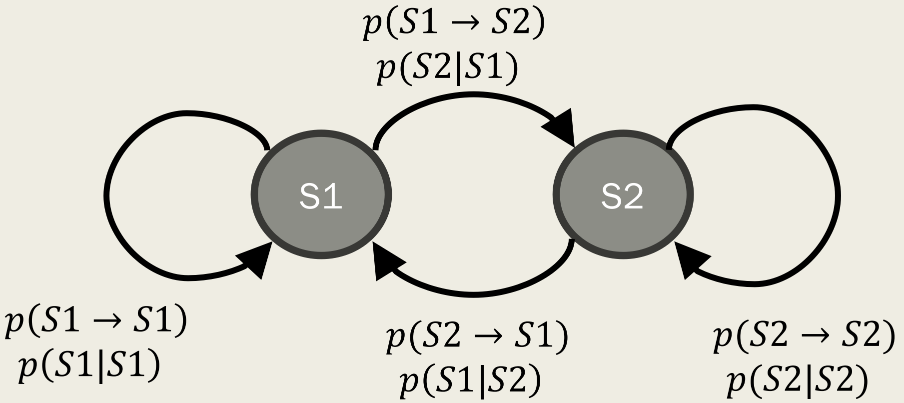
$$= \frac{1}{2}\left[nS^{-1} - \sum\nolimits_{i=1}^{n}\left(x^{(i)} - \mu\right)\left(x^{(i)} - \mu\right)^T\right] = 0$$

$$S^{-1} = \frac{1}{n}\sum\nolimits_{i=1}^{n}\left(x^{(i)} - \mu\right)\left(x^{(i)} - \mu\right)^T = \hat{\Sigma}_{MLE}$$

Markov Chain

- Markov Property: The future state depends only on the current state and not on the sequence of events that preceded it. In other words, "history doesn't matter," only the present does. Mathematically:

$$P(X_{t+1} = s_{t+1} | X_t = s_t, X_{t-1} = s_{t-1}, \dots, X_0 = s_0) = P(X_{t+1} = s_{t+1} | X_t = s_t)$$



Transition probabilities re the probabilities of moving from one state to another.

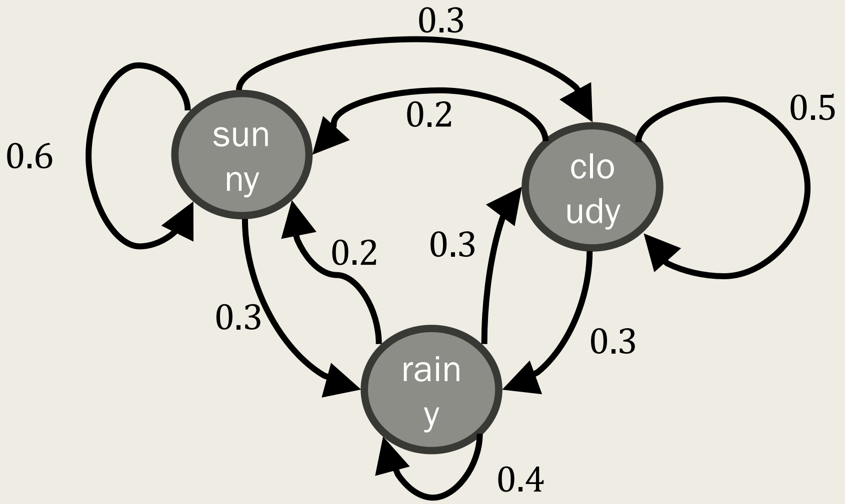
Probability state transition matrix P, where each element P_{ij} represents the probability of transitioning from state i to state j.

Initial State Distribution: This is the probability distribution over the states at the start of the process, often denoted as π. It tells us where we might start in our Markov Chain.

Markov Chain Example: Weather

- Modeling the weather. Suppose we have three possible states: "Sunny," "Cloudy," and "Rainy.” Transition matrix P might look like this:

$$P = \begin{bmatrix} 0.6 & 0.3 & 0.1 \\ 0.2 & 0.5 & 0.3 \\ 0.3 & 0.3 & 0.4 \end{bmatrix}$$



What’s the probability of it being sunny tomorrow given it is rainy today?

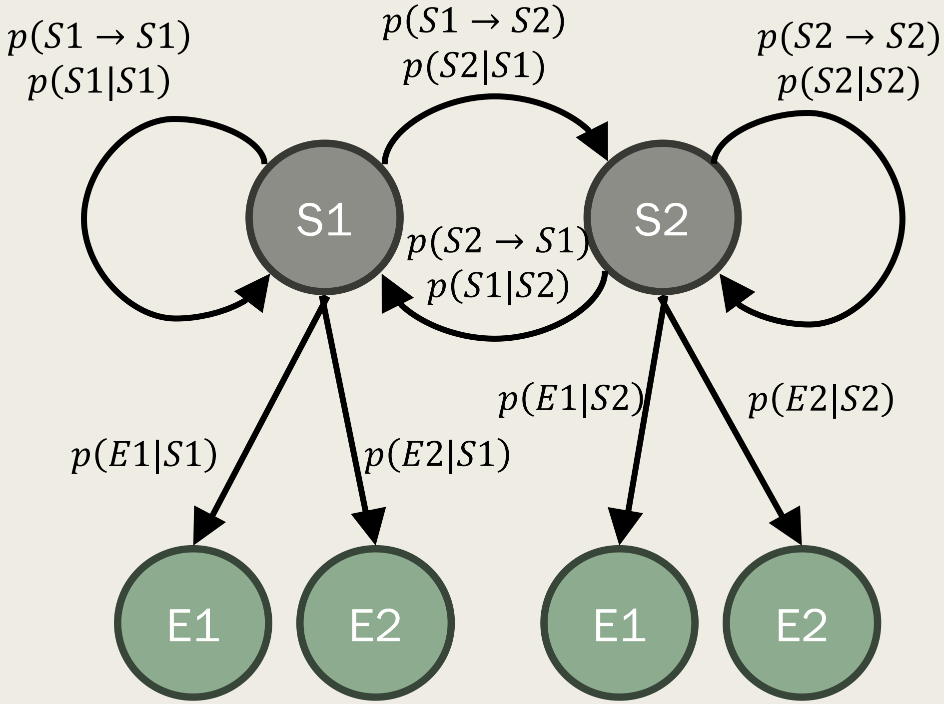
$$P(\text{“}sunny\text{”}|\text{”}rainy\text{”}) = 0.3$$

Over time, a Markov Chain may reach a steady state, where the probabilities of being in each state remain constant from one step to the next.

This is found by solving the equation $\pi A = \pi$, where π is the steady-state distribution.

Hidden Markov Models (HMMs)

A Hidden Markov Model (HMM) is a statistical model where the system being modeled is assumed to follow a Markov process with hidden states. Rather than directly observe the states, we see some evidence that is generated based on the hidden state.



Additional terms to know from Markov chains.

<u>Emissions</u> are what we actually observe (the evidence or state output).

<u>Emissions probabilities</u> are the probabilities of observing a particular outcome given a hidden state.

Hidden Markov Models (HMMs)

An HMM is defined by:

- N hidden states.
- M possible observations.
- A transition matrix A, where A_{ij} is the probability of transitioning from state i to j.
- An emission matrix B, where B_{jk} is the probability of observing k given the hidden state j.
- An initial state distribution π.

HMMs: Weather Example (Extended)

Imagine that we don't directly observe the weather but only whether someone is carrying an umbrella. Thus we have:

- Hidden states: "sunny", "cloudy", or "rainy".
- Observations: "umbrella" or "no umbrella".
- Keep transition probabilities like before, and lets add emission probabilities, such as:
- $P(\text{umbrella}|\text{sunny}) = 0.1$
- $P(\text{umbrella}|\text{cloudy}) = 0.4$
- $P(\text{umbrella}|\text{rainy}) = 0.9$

$$B = \begin{bmatrix} 0.1 & 0.9 \\ 0.4 & 0.6 \\ 0.9 & 0.1 \end{bmatrix} \qquad P = \begin{bmatrix} 0.6 & 0.3 & 0.1 \\ 0.2 & 0.5 & 0.3 \\ 0.3 & 0.3 & 0.4 \end{bmatrix}$$

Find the most likely state-sequence if we observed that in the last three days our subject used an umbrella, then didn't the next day, then did again the last day.

HMMs: Weather Example (Extended)

- We have $P = \begin{bmatrix} 0.6 & 0.3 & 0.1 \\ 0.2 & 0.5 & 0.3 \\ 0.3 & 0.3 & 0.4 \end{bmatrix}$ and $B = \begin{bmatrix} 0.1 & 0.9 \\ 0.4 & 0.6 \\ 0.9 & 0.1 \end{bmatrix}$.
- Observed: umbrella (yes), no umbrella (no), umbrella (yes)
- We want to solve for

$$P(S_1, S_2, S_3 | O_1, O_2, O_3) = \frac{P(O_1, O_2, O_3 | S_1, S_2, S_3) P(S_1, S_2, S_3)}{P(O_1, O_2, O_3)}$$

And since we care for the sequence that maximizes $P(S_1, S_2, S_3 | O_1, O_2, O_3)$, and $P(S_1, S_2, S_3 | O_1, O_2, O_3) \propto P(O_1, O_2, O_3 | S_1, S_2, S_3) P(S_1, S_2, S_3)$, we can just solve for $P(O_1, O_2, O_3 | S_1, S_2, S_3) P(S_1, S_2, S_3)$

However we can solve this by using the probability chain-rule allows us to break down joint probabilities into conditional probabilities:

$$\prod_{i=1}^{n} p(w_i | w_1, \ldots, w_{i-1}, t_1, \ldots, t_n) p(t_i | t_1, \ldots, t_{i-1}) \xrightarrow{\text{By Markov assumption}} \prod_{i=1}^{n} p(O_i | S_i) p(S_i | S_{i-1})$$

HMMs: Weather Example (Extended)

- We have $P = \begin{bmatrix} 0.6 & 0.3 & 0.1 \\ 0.2 & 0.5 & 0.3 \\ 0.3 & 0.3 & 0.4 \end{bmatrix}$ and $B = \begin{bmatrix} 0.1 & 0.9 \\ 0.4 & 0.6 \\ 0.9 & 0.1 \end{bmatrix}$.
- Observed: umbrella (yes), no umbrella (no), umbrella (yes)
- From $\prod_{i=1}^{n} p(O_i|S_i)p(S_i|S_{i-1})$ we get:

$$P(S_1)P(O_1|S_1) * P(S2|S1)P(O2|S2) * P(S3|S2)P(S3|O3)$$

Now calculate this probability sequence for each possibility of S1, S2, and S3 being either sunny, cloudy, or rainy.

The sequence that gives the highest probability is the likeliest sequence.

References

- **Stanford CS229 – Lecture 2:** https://www.youtube.com/watch?v=b0HvwszmqcQ&list=PLoROMvodv4rNH7qL6-efu_q2_bPuy0adh&index=3
- **Stanford CS229 – Lecture 3:** https://www.youtube.com/watch?v=Mi8wnYc1m04&list=PLoROMvodv4rNH7qL6-efu_q2_bPuy0adh&index=3