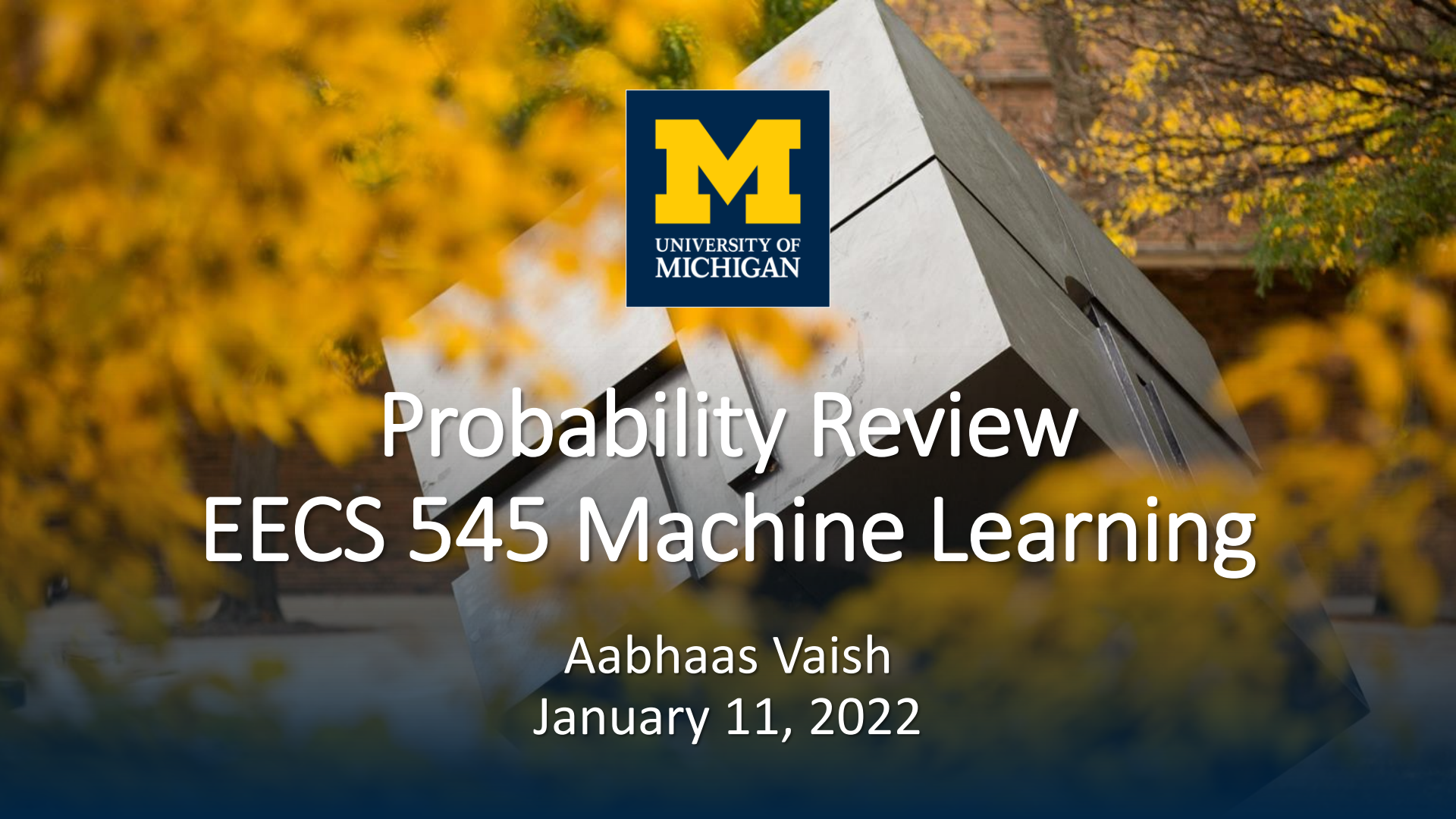


Probability Review
EECS 545 Machine Learning

Aabhaas Vaish

January 11, 2022

Agenda

- Terminology, Law of Total Probability
- Conditional Probability, Independence, Bayes' Rule
- Maximum Likelihood, Maximum A Posteriori
- Expectations and Variances
- MLE and MAP estimates for Gaussian

Slides Adapted from Anthony Liu (EECS 545 WN2021)

What is Probability?

Mathematically, probability is a tool that can be used to model uncertainty. Generally, we are interested in creating models that can measure the degree of uncertainty of an event. For instance:

Source: https://www.freeimages.com/

P(Getting a 2 after rolling a die) = 1/6

There is an uncertainty in obtaining a 2 after rolling a dice. Probability theory allows us to assign a value to this uncertainty.

Terminology

Name	What it is	Common Symbols	What it means
Sample Space	Set	Ω, S	"Possible outcomes."
Event Space	Collection of subsets	$\mathcal{F}, E$	"The things that have probabilities.."
Probability Measure	Measure	P, π	Assigns probabilities to events.
Probability Space	A triple	$(\Omega, \mathcal{F}, P)$	

Remarks: may consider the event space to be the power set of the sample space (for a discrete sample space - more later). e.g., rolling a fair die:

$\Omega = \{1, 2, 3, 4, 5, 6\}$
$\mathcal{F} = 2^{\Omega} = \{\{1\}, \{2\} \ldots \{1, 2\} \ldots \{1, 2, 3\} \ldots \{1, 2, 3, 4, 5, 6\}, \{\}\}$

$P(\{1\}) = P(\{2\}) = \ldots = \frac{1}{6}$ (i.e., a fair die)
$P(\{1, 3, 5\}) = \frac{1}{2}$ (i.e., half chance of odd result)
$P(\{1, 2, 3, 4, 5, 6\}) = 1$ (i.e., result is "almost surely" one of the faces).

Axioms of Probability

In order to maintain consistency, we have the three basic axioms of probability:

1. $P(E) \geq 0$ for any event E
2. $P(\Omega) = 1$ where Ω is the sample space
3. If $A \cap B = \phi$ then $P(A \cup B) = P(A) + P(B)$

Law of Total Probability

Consider events A and B from sample space S. From the law of total probability:

$$P(A) = P(A \cap B) + P(A \cap BC)$$

Question: How?

Now, consider a set of events B_i where they partition the sample space S. We can formulate the law of total probability as:

$$P(A) = \sum_i P(A \cap B_i)$$

Conditional Probability

For events A, B with P(B) > 0, we can find the conditional probability of A given B as:

$$P(A\,|B) = \frac{P(A \cap B)}{P(B)}$$

Note that this is just a re-evaluation of the probability of A after B occurred.

Suppose that we throw a fair die. Consider set A as “result is less than 5” and set B as “result is odd”. What is P(A|B)? What is P(B|A)?

Independence

Two events A, B are said to be independent if $P(A \cap B) = P(A)P(B)$. This can be seen from the fact that $P(A|B) = P(A)$.

Two events A, B are said to be conditionally independent given C when $P(A \cap B|C) = P(A|C)P(B|C)$. From this fact, can we show that $P(B|A,C) = P(B|C)$?

Does independence always imply conditional independence? What about the other way round?

Chain Rule and Independence

Chain Rule: From the definition of conditional probability, we have:

$$p\left(x^{(1)}, \ldots, x^{(N)}\right) = p\left(x^{(N)} | x^{(1)}, \ldots, x^{(N-1)}\right) p\left(x^{(1)}, \ldots, x^{(N-1)}\right)$$

$$p\left(x^{(1)}, \ldots, x^{(N)}\right) = p\left(x^{(N)} | x^{(1)}, \ldots, x^{(N-1)}\right) p\left(x^{(N-1)} | x^{(1)}, \ldots, x^{(N-2)}\right) p\left(x^{(1)}, \ldots, x^{(N-2)}\right)$$

$$= \prod_{i=1}^{N} p\left(x^{(i)} | x^{(1)}, \ldots, x^{(i-1)}\right)$$

Note that random variables $x^{(1)}, \ldots, x^{(N)}$ are independent if and only if:

$$p\left(x^{(1)}, \ldots, x^{(N)}\right) = p(x^{(1)}) \ldots p(x^{(N)})$$

Bayes' Theorem

Using conditional probability, we have:

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

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

Using law of total probability, we can rewrite the denominator as:

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{\sum P(A|B_j)P(B_j)}$$

Note here that B_j must be a partition of the sample space.

Bayes' Theorem inverts the direction of time

Bayes' Theorem Example

- Marie is getting married tomorrow at an outdoor ceremony in the desert. In recent years, it has rained only 5 days each year. Unfortunately, the weatherman is forecasting rain for tomorrow. When it actually rains, the weatherman has forecast rain 90% of the time. When it doesn't rain, he has forecast rain 10% of the time. What is the probability it will rain on the day of Marie's wedding?
- Event *A*: The weatherman has forecast rain.
- Event *B*: It rains.

- We want to know $p(B \mid A)$, the probability it will rain on the day of Marie's wedding, given a forecast for rain by the weatherman. The answer can be determined from Bayes rule:

Bayes' Theorem in ML

Why is Bayes' Theorem so useful in ML? It allows us to compute the posterior of w given data D as:

$$p(w|D) = \frac{p(D|w)p(w)}{p(D)}$$

$$Posterior \propto Likelihood \times Prior$$

Note that the likelihood function, $p(D|w)$, is evaluated for observed data D as a function of w. It's essentially a way to express the probability of observing data D for different w settings.

Maximum Likelihood vs Max A Posteriori

- **Maximum Likelihood Estimate**: Choose parameter setting w that maximizes likelihood function $p(D|w)$

- **Maximum A Posteriori**: Maximize $P(w|D) \propto P(D|w)P(w)$ which is derived from Bayes' Rule. The basic idea here is that, unlike maximum likelihood, we are also considering the fact that the weights have some prior distribution. With this fact, we are trying to find an optimal parameter setting w.

Random Variables: Discrete vs Continuous

A **random variable** is a variable whose outcome is non-deterministic and depends on the result of a random experiment. Generally, random variables are modeled using a PMF or PDF, which (loosely) define the probability of the random variable taking some value.

Discrete RV: Only takes a countable number of values and is generally defined by a Probability Mass Function (PMF) or Cumulative Density Function (CDF).

Continuous RV: Can take infinitely many values since the CDF is continuous everywhere, and is generally defined by Probability Density Function (PDF) or Cumulative Density Function (CDF).

Expectation

If we have X as a random variable with a finite number of outcomes $x^{(1)}, \ldots, x^{(N)}$ occurring with probabilities $p^{(1)}, \ldots, p^{(N)}$, then the expected value of X will be:

$$E(X) = \sum x^{(i)} p^{(i)}$$

Similarly, for the continuous case with PDF $f(x)$, we get:

$$E(X) = \int p(x) f(x) dx$$

- $E[aX] = aE[X]$
- $\mathrm{E[X + Y + Z] = E[X] + E[Y] + E[Z]}$
- $\mathrm{E[XY] = E[X]E[Y]}$ (Independence)

Variance

Variance allows us to measure the spread of the distribution around the expected values (mean). This is measured as:

$$Var(X) = E(X - E[X])^2 = E[X^2] - E[X]^2$$

- $Var(aX) = a^2 Var(X)$
- $Var(X + b) = Var(\mathrm{X})$
- $Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y)$

Covariance and Correlation Coefficient

Covariance measures how one random variable varies with another one. This means that:

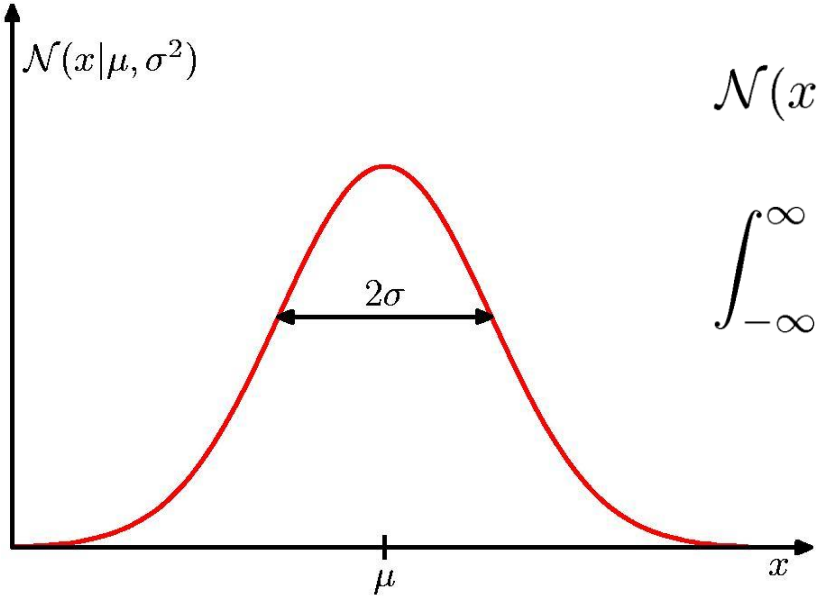
$$Cov(X,Y) = E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y]$$

Finally, we can use this to define the correlation coefficient as:

$$\rho_{XY} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

- $If\ Cov(X,Y) = 0,\ then\ the\ variables\ are\ said\ to\ be\ uncorrelated$
- $Then, we\ have\ E[XY] = E[X]E[Y]$
- $Additionally, in\ this\ case, we\ have\ Var(X+Y) = Var(X) + Var(Y)$

Gaussian Distribution



$$\mathcal{N}\left(x|\mu,\sigma^2\right) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

$$\mathcal{N}(x|\mu,\sigma^2) > 0$$

$$\int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu,\sigma^2\right)\,\mathrm{d}x = 1$$

MLE for Gaussian

Consider a dataset of samples from a Gaussian RV, $D = \{x^{(1)}, \ldots, x^{(N)}\}$ and we know that the variance of the data is σ^2. Find the MLE estimate of μ which maximizes probability $p(D|\mu)$.

$$\boldsymbol{\mu_{MLE}} = \frac{\mathbf{1}}{\boldsymbol{N}} \sum \boldsymbol{x^{(N)}}$$

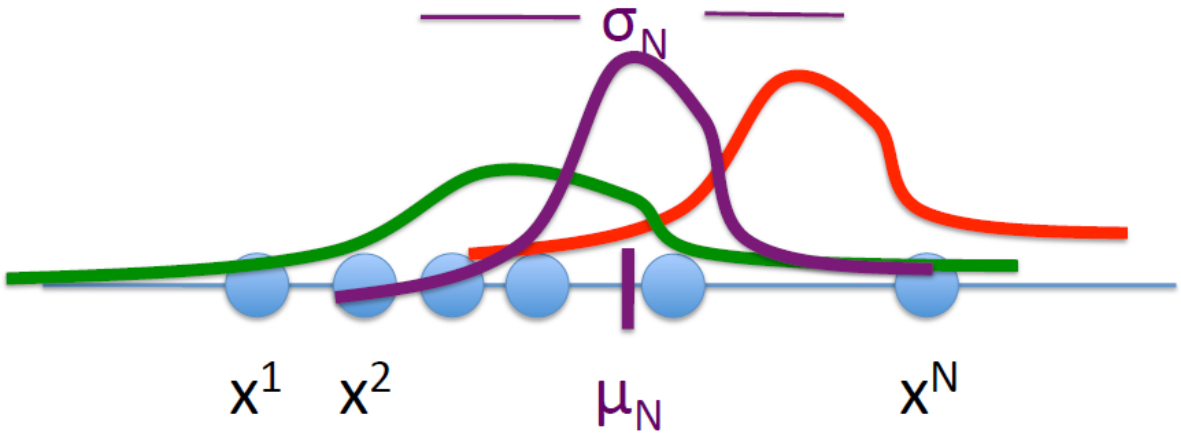
MLE for Gaussian – For Annotation

MAP for Gaussian

Consider a dataset of samples from a Gaussian RV, $D = \{x^{(1)}, \ldots, x^{(N)}\}$ and we know that the variance of the data is σ^2. We also assume that μ is distributed normally with mean μ_0 and variance σ_0^2. Find the MAP estimate of μ which maximizes probability $p(\mu|D)$.

MAP for Gaussian – For Annotation

MAP vs MLE



Prior belief

Maximum Likelihood

Posterior Distribution

Multivariate Gaussian Distribution

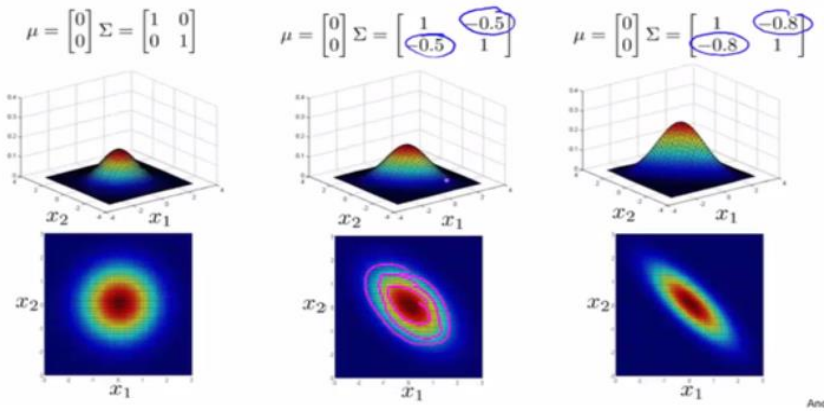
We can compute the PDF as follows:

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

μ: Mean vector (d by 1)

Σ: Covariance matrix (d by d)

$|\Sigma|$: Matrix determinant



Credits: Andrew Ng

Thank You!