EECS 545 - Machine Learning
Review Session 3: Probability

01/21/2020

Geunseob (GS) Oh

Ph.D. Candidate

University of Michigan

Credits: Jim Boerkoel(UofM), Jeff Howbert(CMU), Shenlong Wang(UofToronto), and Bishop's PRML

Outline

- Terminology, Law of Total Probability
- Conditional probability, Independence, Bayes' rule
- Maximum likelihood, Maximum a posteriori
- MLE and MAP estimation for 1D Gaussian
- Expectations and Variances
- Distributions

Probability

- The world is full of uncertainty

Probability of Wolverines beat Spartans in 2021? $P(W_{2021})$

Probability of rain tomorrow? $P(R_{tom})$

$P(W_{2021}|W_{2020})$

We beat Spartans (2020). Probability of Wolverines beat Spartans in 2021?

The weather is rainy today. Probability of rain tomorrow? $P(R_{tom}|R_{today})$

- Probability is a tool to represent uncertainty
- We build models to estimate the uncertainty using probability

$P_{model}(W_{2021})$

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).

Law of Total Probability

- $P(A) \geq 0,\ \forall A \in F$
- $P(\Omega) = 1$
- Law of total probability

$$P(A) = P(A \cap B) + P(A \cap B^C)$$

$$P(A) = \sum_i P(A \cap B_i) \qquad \text{Discrete } B_i$$

$$P(A) = \int P(A \cap B_i) dB_i \qquad \text{Continuous } B_i$$

Conditional Probability

For events $A, B \in \mathcal{F}$ with $P(B) > 0$, we may write the **conditional probability of A given B**:

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

$P(A, B)$ Joint probability of A and B

$P(A)$ Marginal probability of A

probability of A given B is true

Suppose we throw a fair die:

$\Omega = \{1, 2, 3, 4, 5, 6\}$, $\mathcal{F} = 2^{\Omega}$, $P(\{i\}) = \frac{1}{6}$, $i = 1 \ldots 6$

$A = \{1, 2, 3, 4\}$ i.e., "result is less than 5,"

$B = \{1, 3, 5\}$ i.e., "result is odd."

What is the probability of A given B?

Probability of B given A?

Conditional Probability

Suppose we throw a fair die:
$\Omega = \{1,2,3,4,5,6\},\ \mathcal{F} = 2^{\Omega},\ P(\{i\}) = \frac{1}{6},\ i = 1 \ldots 6$
$A = \{1,2,3,4\}$ i.e., "result is less than 5,"
$B = \{1,3,5\}$ i.e., "result is odd."

What is the probability of A given B?
Probability of B given A?

$$\begin{aligned} P(A) &= \frac{2}{3} \\ P(B) &= \frac{1}{2} \end{aligned}$$

$$\begin{aligned} P(A|B) &= \frac{P(A \cap B)}{P(B)} \\ &= \frac{P(\{1,3\})}{P(B)} \\ &= \frac{2}{3} \end{aligned}$$

$$\begin{aligned} P(B|A) &= \frac{P(A \cap B)}{P(A)} \\ &= \frac{1}{2} \end{aligned}$$

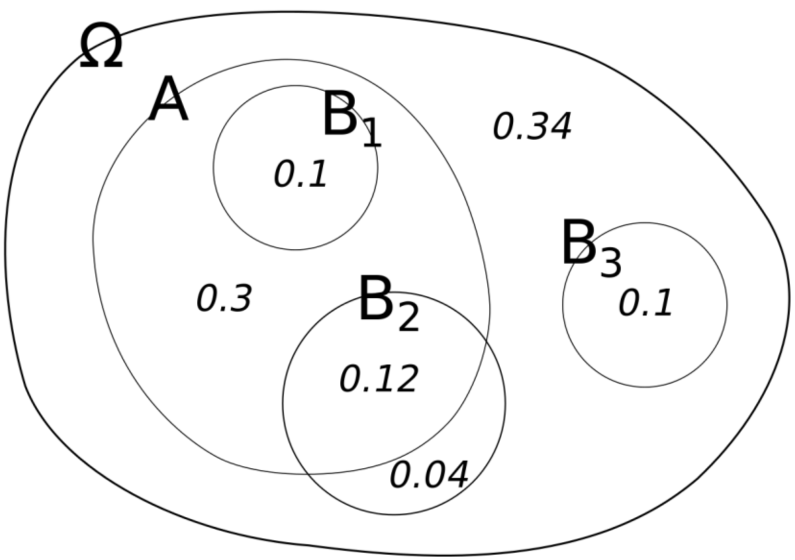
Conditional Probability

For events $A, B \in \mathcal{F}$ with $P(B) > 0$, we may write the **conditional probability of A given B**:

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

$P(A,B)$ Joint probability of A and B

$P(A)$ Marginal probability of A



From Wikipedia

$P(A|B_1) = 1$

$P(A|B_2) = 0.12 \div (0.12 + 0.04) = 0.75$

$P(A|B_3) = 0$ (disjoint)

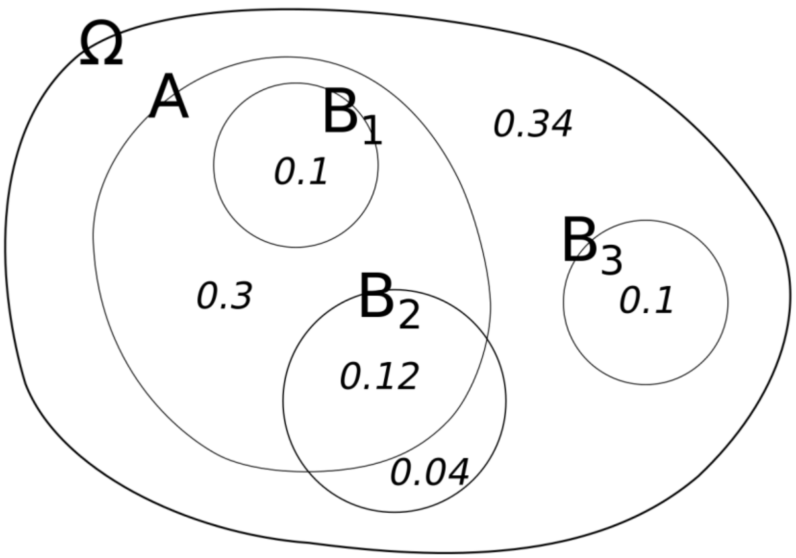
$B_4 = (B_1 \cup B_2 \cup B_3)^c$

$P(A, B_4) = 0.3$

Conditional Probability w Law of Total Prob

For events $A, B \in \mathcal{F}$ with $P(B) > 0$, we may write the **conditional probability of A given B**:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \qquad \& \qquad P(A) = \sum_i P(A \cap B_i)$$



P(A) (The unconditional probability)

$$= \overset{0.1}{P(A, B_1)} + \overset{0.12}{P(A, B_2)} + \overset{0}{P(A, B_3)} + \overset{0.3}{P(A, B_4)}$$

$$= \overset{1 * 0.1}{P(A|B_1)P(B_1)} + \overset{0.75*0.16}{P(A|B_2)P(B_2)}$$

$$+ \overset{0}{P(A|B_3)P(B_3)} + \overset{0.3/\, P(B_4)*\, P(B_4)}{P(A|B_4)P(B_4)}$$

= 0.52

From Wikipedia

Independence

Two events A, B are called **independent** if $P(A \cap B) = P(A)P(B)$.

When $P(A) > 0$ this may be written $P(B|A) = P(B)$ (why?)
e.g., rolling two dice, flipping n coins etc.

Two events A, B are called **conditionally independent given** C when $P(A \cap B|C) = P(A|C)P(B|C)$.

When $P(A) > 0$ we may write $P(B|A, C) = P(B|C)$
e.g., "the weather tomorrow is independent of the weather yesterday, knowing the weather today."

Independence -> Conditional Independence ?

Conditional Independence -> Independence ?

Conditional Probability & Independence

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

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

If A, B are conditionally independent given C:

$$P(A,B|C) = P(A|C)P(B|C)$$

Chain Rule and Independence

Chain rule: From the definition of conditional probabilities, one can show that

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

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

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

Random variables $\mathrm{x}^{(1)}, \dots, \mathrm{x}^{(N)}$ are **independent** if and only if

$$p\left(\mathrm{x}^{(1)}, \dots, \mathrm{x}^{(N)}\right) = p\left(\mathrm{x}^{(1)}\right) p\left(\mathrm{x}^{(2)}\right) \cdots p\left(\mathrm{x}^{(N)}\right)$$

Lecture 3: Linear Regression Revisited

Linear regression modeled using gaussian distribution:

$$y^{(n)} = \mathbf{w}^T \phi(\mathbf{x}^{(n)}) + \epsilon$$

$$p(y^{(n)} | \phi(\mathbf{x}^{(n)}), \mathbf{w}, \beta) = \mathcal{N}(y^{(n)} | \mathbf{w}^T \phi(\mathbf{x}^{(n)}), \beta^{-1})$$

We quickly went over:

$$\log p(y^{(1)}, y^{(2)}, \ldots, y^{(N)} | \boldsymbol{\Phi}, \mathbf{w}, \beta) = \log \prod_{n=1}^{N} \mathcal{N}(y^{(n)} | \mathbf{w}^T \phi(\mathbf{x}^{(n)}), \beta^{-1})$$

Detailed derivation:

$$p\left(y^{(1)}, \ldots, y^{(N)} \middle| \mathrm{x}^{(1)}, \ldots, \mathrm{x}^{(N)}, \mathbf{w}, \beta\right)$$

Using Chain Rule

$$= p(y^{(N)} | y^{(1)}, \ldots, y^{(N-1)}, \mathrm{x}^{(1)}, \ldots, \mathrm{x}^{(N)}, \mathbf{w}, \beta)\, p(y^{(1)}, \ldots, y^{(N-1)} | \mathrm{x}^{(1)}, \ldots, \mathrm{x}^{(N)}, \mathbf{w}, \beta)$$

$$= \prod_{n=1}^{N} p(y^{(n)} | y^{(1)}, \ldots, y^{(n-1)}, \mathrm{x}^{(1)}, \ldots, \mathrm{x}^{(N)}, \mathbf{w}, \beta)$$

$$= \prod_{n=1}^{N} p\left(y^{(n)} \middle| \mathrm{x}^{(n)}, \mathbf{w}, \beta\right) = \prod_{n=1}^{N} N\left(y^{(n)} \middle| \mathbf{w}^T \phi(\mathrm{x}^{(n)}), \beta^{-1}\right)$$

Bayes' Theorem

Using the chain rule we may see:

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

Rearranging this yields **Bayes' rule**:

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

Often this is written as:

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

Where B_i are a partition of Ω (note the bottom is just the law of total probability).

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, 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 know:
  - $p(B) = 5 / 365 = 0.0137$ [It rains 5 days out of the year.]
  - $p(\text{ not } B) = 360 / 365 = 0.9863$
  - $p(A \mid B) = 0.9$ [When it rains, the weatherman has forecast rain 90% of the time.]
  - $p(A \mid \text{not } B) = 0.1$ [When it does not rain, the weatherman has forecast rain 10% of the time.]

Bayes' Theorem, Example

- We know:
  - $p(B) = 5 / 365 = 0.0137$ [It rains 5 days out of the year.]
  - $p(\text{not } B) = 360 / 365 = 0.9863$
  - $p(A \mid B) = 0.9$ [When it rains, the weatherman has forecast rain 90% of the time.]
  - $p(A \mid \text{not } B) = 0.1$ [When it does not rain, the weatherman has forecast rain 10% of the time.]

What we would like to compute using Bayes' Rule:

1. $p(B \mid A) = p(A \mid B) \cdot p(B) / p(A)$

Obtain $P(A)$ using Law of Total Probability:

2. $p(A) = p(A \mid B) \cdot p(B) + p(A \mid \text{not } B) \cdot p(\text{not } B) =$
   $(0.9)(0.014) + (0.1)(0.986) = 0.111$

3. $p(B \mid A) = (0.9)(0.0137) / 0.111 = 0.111$

Bayes' Theorem in Learning

- Why is Bayes' so useful in learning? Allows us to compute the posterior of w given data D:

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

Prior

Posterior

Likelihood

$$p(D) = \int p(D|w)p(w)dw$$

- Bayes' rule in words: posterior ∝ likelihood × prior

$$p(\mathbf{w}|D) \propto p(D|\mathbf{w})p(\mathbf{w})$$

- The likelihood function, $p(D|w)$, is evaluated for observed data D as a function of w. It expresses how probable the observed data set is for various parameter settings w.

Maximum Likelihood vs Maximum A Posteriori

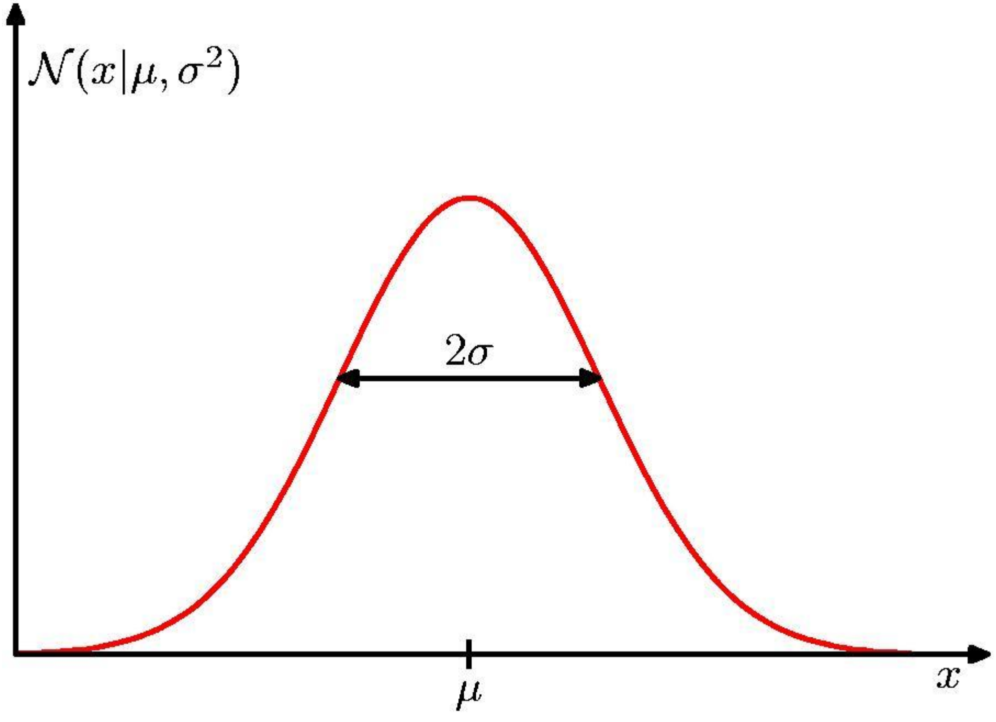
- Maximum likelihood:
  - choose parameter setting w that maximizes likelihood function $p(D|w)$.
  - Choose the value of w that maximizes the probability of observed data.

- Cf. MAP (Maximum a posteriori) estimation
  - Equivalent to maximizing $P(w|D) \propto P(D|w)P(w)$
  - Can compute this using Bayes rule!
  - This will be covered in later lectures

Frequentism and Bayesianism: What's the Big Deal?
https://www.youtube.com/watch?v=KhAUfqhLakw

Gaussian Distribution

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



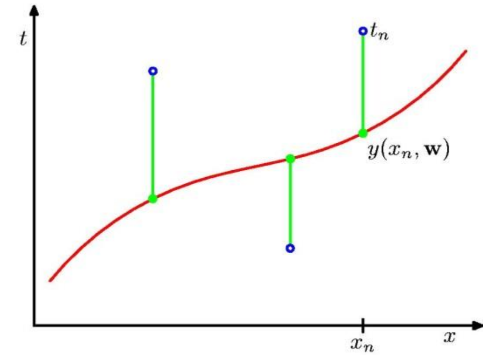
$$\mathcal{N}(x|\mu,\sigma^2) > 0$$

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

Maximum Likelihood Estimation (MLE) for Linear Regression

- Assume a stochastic model:

$$y^{(n)} = \mathbf{w}^T \phi(\mathbf{x}^{(n)}) + \epsilon \text{ where } \epsilon \sim \mathcal{N}(0, \beta^{-1})$$



- This gives a likelihood function:

$$p(y^{(n)}|\phi(\mathbf{x}^{(n)}), \mathbf{w}, \beta) = \mathcal{N}(y^{(n)}|\mathbf{w}^T\phi(\mathbf{x}^{(n)}), \beta^{-1})$$

- With input matrix $\mathbf{\Phi}$ and output matrix $\mathbf{y}$, the data likelihood is:

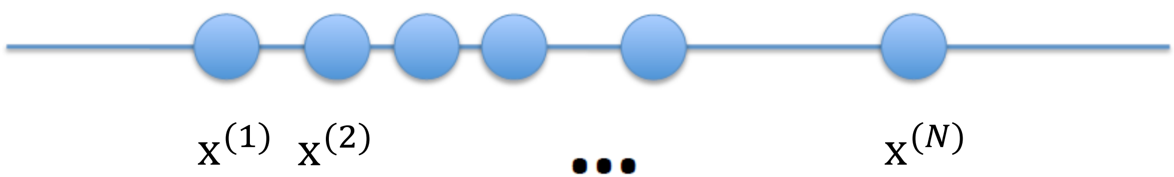
$$p(\mathbf{y}|\mathbf{\Phi}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(y^{(n)}|\mathbf{w}^T\phi(\mathbf{x}^{(n)}), \beta^{-1})$$

$p(D|\mathbf{w})$ likelihood

$$\log p(\mathbf{y}|\mathbf{\Phi}, \mathbf{w}, \beta) = N \log \beta - \frac{N}{2} \log 2\pi - \beta E_D(\mathbf{w})$$

MLE for 1D Gaussian

- Problem: Suppose we are given a data set of samples of a Gaussian random variable X, $D = \{x^{(1)}, \ldots, x^{(N)}\}$ and told that the variance of the data is σ^2



- What we want to get:

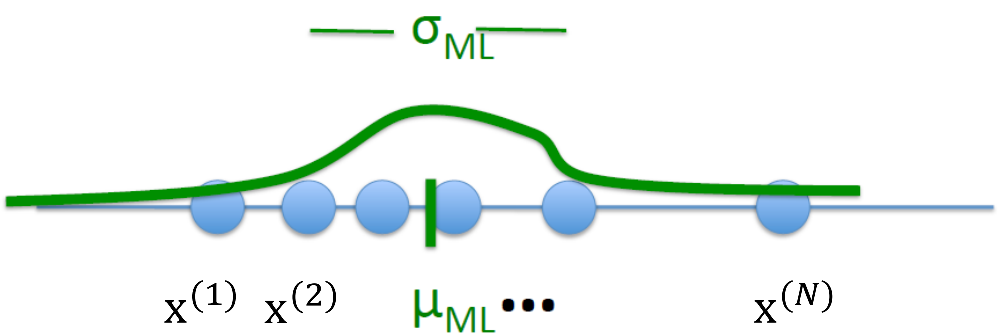
  μ that best fits the data points

  μ that maximizes the probability $p(D|\mu)$

MLE for 1D Gaussian

- What we want to get:

  μ that maximizes the probability $p(D|\mu)$



Maximum Likelihood

$$p(D|\mu) = p(\mathrm{x}^{(1)}, \mathrm{x}^{(2)}, \ldots, \mathrm{x}^{(N)}|\mu)$$

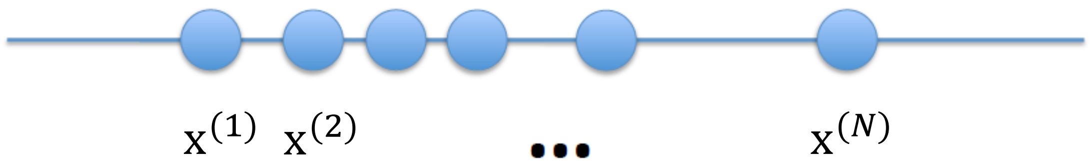
$$= p\left(\mathrm{x}^{(1)}\middle|\mu\right)p(\mathrm{x}^{(2)}|\mu), \ldots, p(\mathrm{x}^{(N)}|\mu)$$

$$\log(p(D|\mu)) = \sum \log(p(\mathrm{x}^{(n)}|\mu))$$

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

MAP for 1D Gaussian

- Problem: Suppose we are given a data set of samples of a Gaussian random variable X, $D = \{x^{(1)}, \dots, x^{(N)}\}$ and told that the variance of the data is σ^2



- What we want to get:

  $p(\mu|D)$, The distribution of μ after observing D

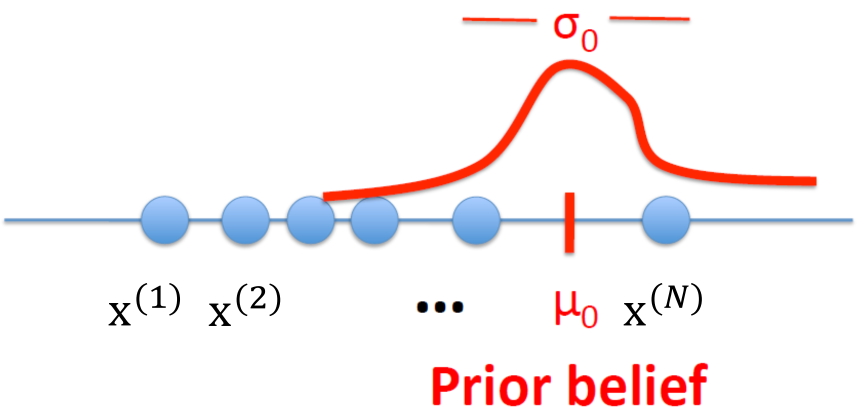
- Let's say we believe that μ is a random variable distributed normally with mean μ_0 variance σ_0^2

$$\boldsymbol{p(\mu) = N(\mu_0, \sigma_0^2)}$$

MAP for 1D Gaussian

$$p(\mu) = N(\mu_0, \sigma_0^2)$$

$$= \frac{1}{\sqrt{2\pi}\sigma_0} \exp(-\frac{(\mu - \mu_0)^2}{2\sigma_0^2})$$



- $p(\mu)$ is the prior probability of μ
- What we want to get: $p(\mu|D) = \frac{p(D|\mu)p(\mu)}{p(D)}$
- Since D is from a 1D Gaussian,

$$p(D|\mu) = p(\mathrm{x}^{(1)}, \mathrm{x}^{(2)}, \dots, \mathrm{x}^{(N)}|\mu)$$

$$= \prod p(\mathrm{x}^{(n)}|\mu) = \prod \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(\mathrm{x}^{(n)} - \mu)^2}{2\sigma^2})$$

MAP for 1D Gaussian

- What we want to get: $p(\mu|D) = \dfrac{p(D|\mu)p(\mu)}{p(D)}$ Constant

- $p(D|\mu)p(\mu) = \prod[\frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{(\mathrm{x}^{(n)}-\mu)^2}{2\sigma^2})]\frac{1}{\sqrt{2\pi}\sigma_0}\exp(-\frac{(\mu-\mu_0)^2}{2\sigma_0^2})$

"The product of two Gaussian pdfs = A bivariate Gaussian pdf"

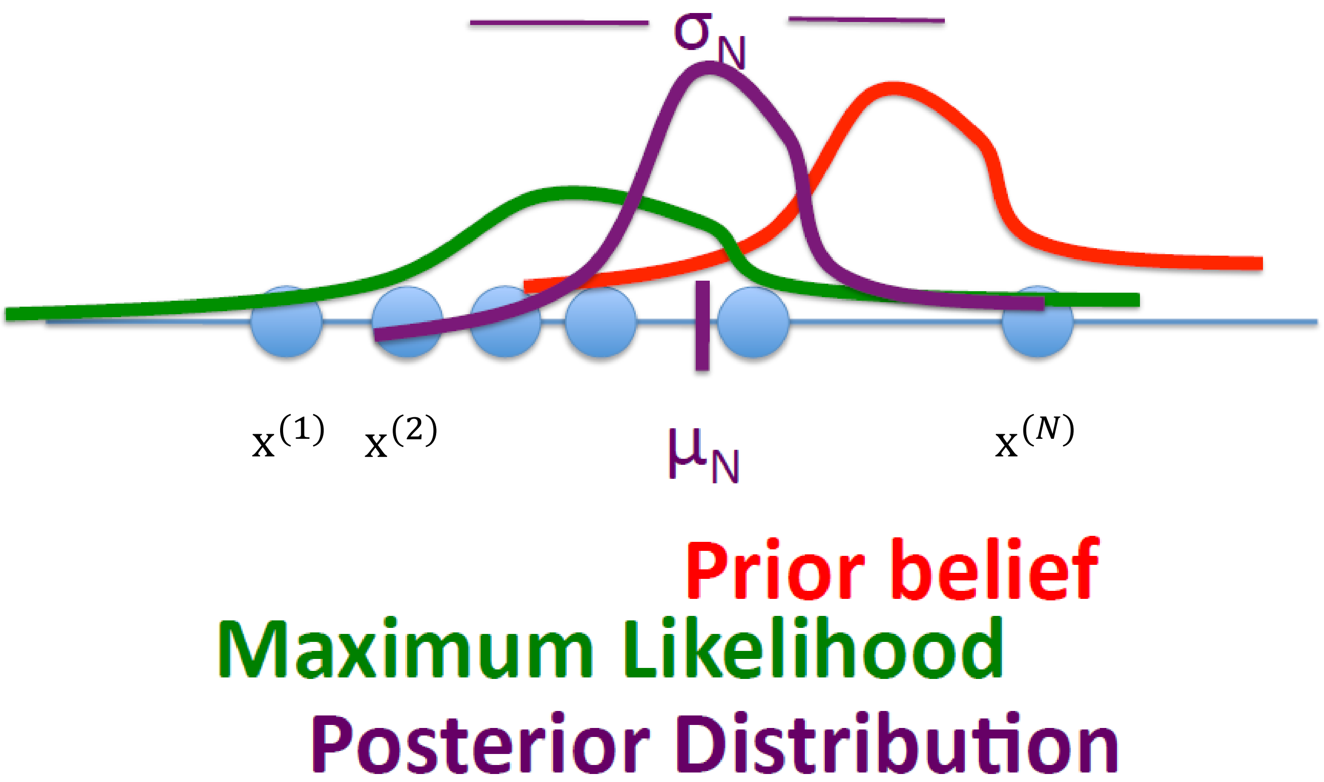
- $p(\mu|D) = \mathrm{N}(\mu|\mu_N, \sigma_N)$

where $$\mu_N = \frac{\sigma^2}{N{\sigma_0}^2 + \sigma^2}\mu_0 + \frac{N{\sigma_0}^2}{N{\sigma_0}^2 + \sigma^2}\mu_{ML}$$

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

$$\frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}$$

MAP for 1D Gaussian



Random Variables: Discrete vs Continuous

- Discrete RV: only takes a countable number of values

  Distribution defined by probability mass function (PMF)

  or cumulative density function (CDF)

  Marginalization: $p(x) = \sum_y p(x, y)$

- Continuous RV: takes infinitely many values (its CDF is continuous everywhere)

  Distribution defined by probability density function (PDF)

  or cumulative density function (CDF)

  Marginalization: $p(x) = \int_y p(x, y) dy$

Expectations

- Let X be a random variable with a finite number of outcomes $x^{(1)}, \ldots, x^{(N)}$ occurring with probabilities $p^{(1)}, \ldots, p^{(N)}$, then the expectation of X:

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

- The expected value of the function $f(x)$ given that x has a probability density function p(x):

[Discrete] $$\mathbb{E}[f] = \sum_x p(x) f(x)$$

Q. What is the expected value of a roll of a fair die?

[Continuous] $$\mathbb{E}[f] = \int p(x) f(x) \, \mathrm{d}x$$

Q. What is the expected value of $f(x) = 1$ where x is drawn from standard normal distribution ?

Variance

- Variance: measures how far a set of (random) numbers are spread out from the expected value

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

Q. Variance of a coin toss?

- $Var[a] = 0$ for any constant $a \in \mathbb{R}$.
- $Var[af(X)] = a^2 Var[f(X)]$ for any constant $a \in \mathbb{R}$.

- $E[a] = a$ for any constant $a \in \mathbb{R}$.
- $E[af(X)] = aE[f(X)]$ for any constant $a \in \mathbb{R}$.
- (Linearity of Expectation) $E[f(X) + g(X)] = E[f(X)] + E[g(X)]$.

Expectation and Covariance Multi-variable Distribution

Expectation

[Discrete] $$E[g(X,Y)] \triangleq \sum_{x \in Val(X)} \sum_{y \in Val(Y)} g(x,y) p_{XY}(x,y).$$

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

Covariance

$$Cov[X,Y] \triangleq E[(X - E[X])(Y - E[Y])]$$

$$\begin{aligned} Cov[X,Y] &= E[(X - E[X])(Y - E[Y])] \\ &= E[XY - XE[Y] - YE[X] + E[X]E[Y]] \\ &= E[XY] - E[X]E[Y] - E[Y]E[X] + E[X]E[Y]] \\ &= E[XY] - E[X]E[Y]. \end{aligned}$$

Expectation and Covariance Multi-variable Distribution

- If X and Y are independent, then $Cov[X, Y] = 0$.
- If X and Y are independent, then $E[f(X)g(Y)] = E[f(X)]E[g(Y)]$.

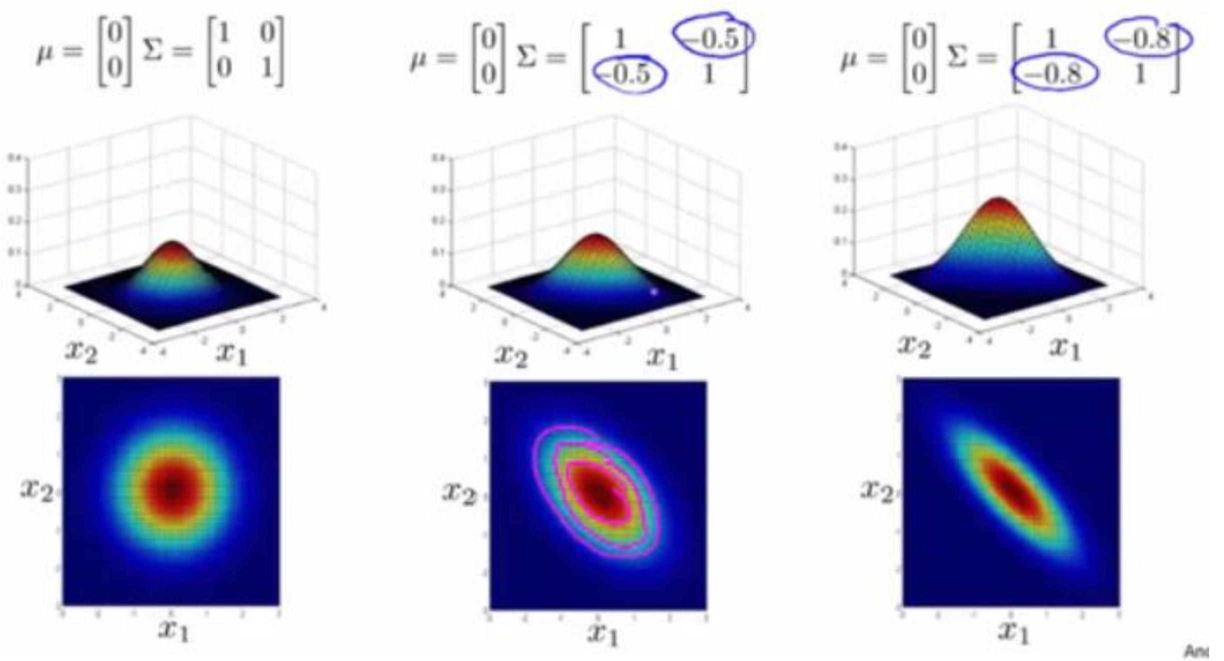
- (Linearity of expectation) $E[f(X, Y) + g(X, Y)] = E[f(X, Y)] + E[g(X, Y)]$.
- $Var[X + Y] = Var[X] + Var[Y] + 2Cov[X, Y]$.

Multivariate Gaussian Distribution

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

μ: Mean vector (d by 1)
Σ: Covariance matrix (d by d)
$|\Sigma|$: Matrix determinant

Bi-variate (2D) Gaussian:
(Credit to Andrew Ng)



Common Discrete Random Variables

- $X \sim Bernoulli(p)$ (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.

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

Common Continuous Random Variables

- $X \sim Uniform(a, b)$ (where $a < b$): equal probability density to every value between a and b on the real line.

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

- $X \sim Exponential(\lambda)$ (where $\lambda > 0$): decaying probability density over the nonnegative reals.

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

- $X \sim Normal(\mu, \sigma^2)$: also known as the Gaussian distribution

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

Properties of Common Distributions

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 $0 \leq k \leq n$	np	npq
$Geometric(p)$	$p(1-p)^{k-1}$ for $k = 1, 2, \ldots$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
$Poisson(\lambda)$	$e^{-\lambda}\lambda^x/x!$ for $k = 1, 2, \ldots$	λ	λ
$Uniform(a,b)$	$\frac{1}{b-a}$ $\forall x \in (a,b)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
$Gaussian(\mu,\sigma^2)$	$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	μ	σ^2
$Exponential(\lambda)$	$\lambda e^{-\lambda x}$ $x \geq 0, \lambda > 0$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$