CEE 616: Probabilistic Machine Learning
Lecture 1a: Foundations: Probability

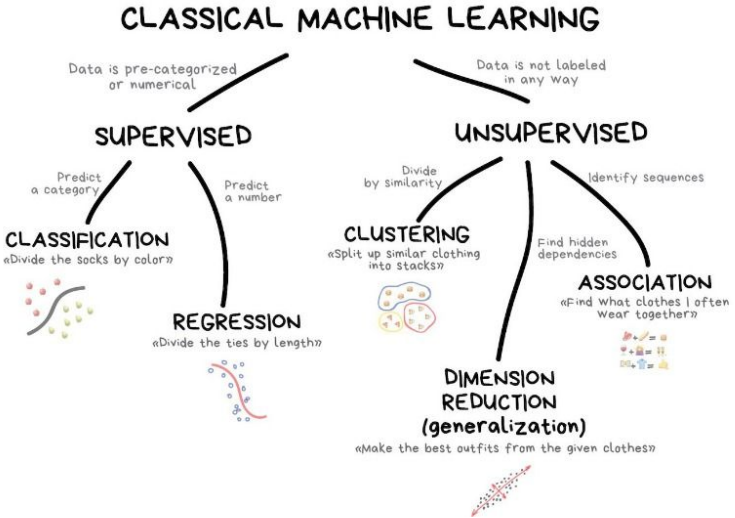
Jimi Oke

UMass Amherst
College of Engineering

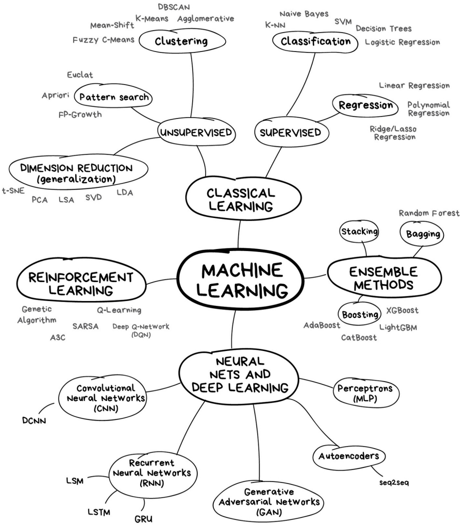
Feb 8, 2023

Outline

What is machine learning?



Machine learning—alternate illustration



Source: `https://i.vas3k.ru/7vx.jpg`

Machine learning flow

The "learning" refers to the search for **optimal parameters** as a function of the data.

- Inputs (data, domain knowledge/human)
- Learning (computer/algorithm)
- Outputs (predictions, information/inference)

Supervised vs. unsupervised learning

7

Supervised learning

Goal: fit a model characterizing relationship between predictor(s) X and response(s) Y (i.e. known outputs)

- regression (linear, nonlinear, logistic, etc)
- boosting/bagging/random forests
- support vector machines

Unsupervised learning

Goal: infer relationships between/among variables or observations (outputs/target unknown)

- dimensionality reduction (principal components, factor analysis)
- cluster analysis

- Semi-supervised learning occurs when responses are available for a subset of the observations

Notation

Symbol	**Meaning**
n	number of observations (distinct data points)
p	number of variables
x_{ij}	value of jth variable for ith observation

So, we can write the $n \times p$ matrix $\boldsymbol{X}$ as:

$$\boldsymbol{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

where the rows of $\boldsymbol{X}$ are: $x_1, x_2, \ldots, x_n$
and the columns are written $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p$

Notation (cont.)

We will denote $\boldsymbol{y}$ as the *response* variable vector.

Summary of notation conventions

- Scalar: lower case italic (e.g. b)
- Vector: lower case bold (e.g $\boldsymbol{x}_j \in \mathbb{R}^n$), except for feature vectors of length p)
- Matrix: upper case bold (e.g. $\boldsymbol{X} \in \mathbb{R}^{n \times p}1$)
- Random variable: upper case italic (e.g. $Y \sim \mathcal{N}(\mu, \sigma)$)

Learning framework

Given a set of inputs X_j ($j \in \{1, \ldots, p\}$) and a given output Y, **"learning"** refers to the techniques used in estimating the functional relationship between X_i and Y for the purposes of *prediction* and *inference*.

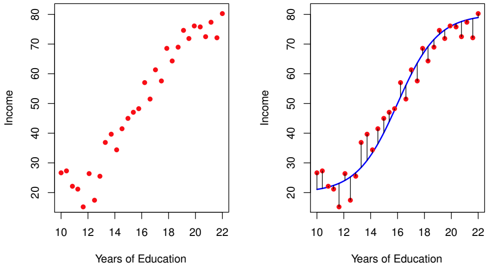


Figure: Estimating the functional relationship between income and educational attainment in a data set

Model equation

$$Y = f(X) + \epsilon \tag{1}$$

where f is an unknown function and ϵ is the random error (independent of X with zero mean)

Prediction

We predict Y using:

$$\hat{Y} = \hat{f}(X) \tag{2}$$

where $\hat{f}$ is the estimate of f and $\hat{Y}$ is the predicted value of Y.

Reducible and irreducible error

The prediction accuracy depends on *reducible error* and *irreducible error* (noise—intrinsic variability in the data)

$$\begin{aligned} E\left[(Y - \hat{Y})^2\right] &= E\left[f(X) + \epsilon - \hat{f}(X))^2\right] \\ &= \left[f(X) - \hat{f}(X)\right]^2 + \mathrm{Var}(\epsilon) \end{aligned} \tag{3}$$

Inference

This refers to the process of determining the nature of the relationship between the inputs (X) and outputs (Y).
In other words, if $Y = f(X)$, then what is f?

Questions relating to inference

- What is the elasticity[a] of a certain input in relation to an output?
- What are the important *predictors* of a certain outcome?
- What is the correlation between X and Y?

[a] Can be defined as the percentage change in Y for a 1% change in X.

Parametric methods

These methods require an assumption of the structure of the relationship between X and Y.

Step 1: assume functional form (e.g. linearity in coefficients):

$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots \tag{4}$$

Step 2: fit model, i.e. *estimate* the parameters/coefficients:

$$Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots \tag{5}$$

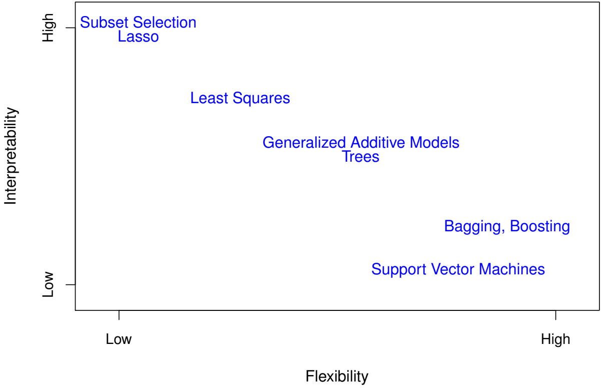
The estimation procedure can be a method of choice, e.g. OLS (ordinary least squares), WLS (weighted least squares), etc.

Non-parametric methods

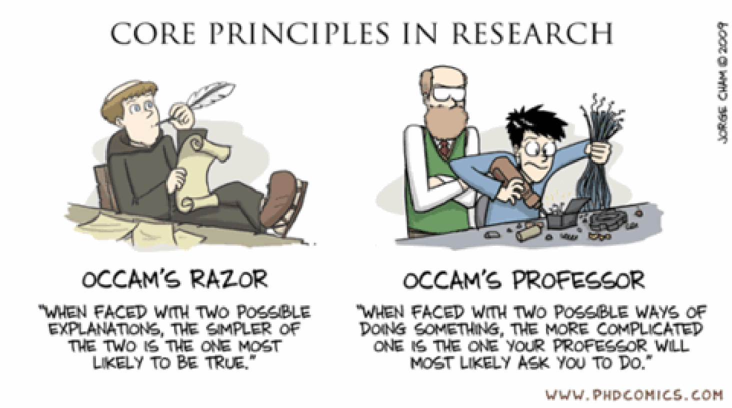
- The assumption of linearity is a strong one and may result in a poor fit if f is very different from $\hat{f}$.
- Non-parametric methods allow flexible functional forms (although the danger of overfitting is real).
- For accuracy, however, non-parametric models require many more observations compared to the parametric case.

Tradeoff between accuracy and interpretability

A simpler model is more interpretable in its parameters. A highly complicated model may operate more like a blackbox.



Occam's razor



Theory of probability

- Three axioms of probability:

$$
\begin{aligned}
P(E) \geq 0 \quad \text{and} \quad P(E) \leq 1 \quad & \text{for given event } E \\
P(S) = 1 & \\
P(E_1 \cup E_2 \cup \cdots \cup E_n) = P(E_1) + P(E_2) + \cdots + P(E_n) \quad & \text{(Mutually exclusive)}
\end{aligned}
$$

- Addition rule: $P(A \cup B) = P(A) + P(B) - P(AB)$
  - For mutually exclusive events: $P(A \cup B) = P(A) + P(B)$ (Axiom 3)
- Counting events:
  - Fundamental principle of counting: number of outcomes for $1, \ldots, k$ events, each with $n_1, \ldots, n_k$ possibilities is $n_1 \times \cdots \times n_k$
  - Permutations (arrangements) of n objects: $n! = n(n-1)(n-2)\cdots(2)(1)$
  - Permutations of a subset of k items chosen from set of n items: $n!/(n-k)!$
  - Combinations (distinct; order not important) of group of k items chosen from set of n items: $n!/(k!(n-k)!)$

Conditional probability

- Conditional probability:
$$P(A|B) = \frac{P(AB)}{P(B)} \tag{6}$$

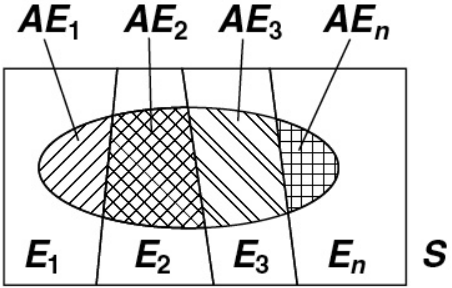
- Independent events:
$$P(AB) = P(A)P(B) \tag{7}$$

- Generally, the joint probability (intersection) of any number of independent events is the product of their individual probabilities:
$$P(E_1 \cap E_2 \cap \cdots \cap E_n) = P(E_1)P(E_2)\cdots P(E_n) \tag{8}$$

- Multiplication rule:
$$P(AB) = P(A|B)P(B) = P(B|A)P(A) \tag{9}$$

Total probability

Useful in situations where the probability of an event cannot be directly determined but its conditional probabilities are known.



$$P(A) = P(AE_1) + P(AE_2)$$

$$+P(AE_3) + \cdots + P(AE_n)$$

Note that:
$P(AE_1) = P(A|E_1)P(E_1)$,
etc.

Theorem of total probability

The probability of an event A conditioned on the mutually exclusive and collectively exhaustive events $E_1, E_2, \ldots, E_n$ is given by

$$P(A) = P(A|E_1)P(E_1) + P(A|E_2)P(E_2) + \cdots + P(A|E_n)P(E_n) \tag{10}$$

Derivation of Bayes' theorem

Recall from the multiplication rule that:

$$P(AB) = P(A|B)P(B) \tag{11}$$

Equivalently:

$$P(AB) = P(B|A)P(A) \tag{12}$$

We combine both equations to obtain:

$$P(A|B)P(B) = P(B|A)P(A) \tag{13}$$

Then, we obtain the **inverse probability** of the conditioning event:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} \tag{14}$$

Bayes' theorem

Bayes' Theorem allows for the computation of an inverse probability, e.g. given $P(A|B)$, can we find $P(B|A)$?

$$P(E_i|A) = \frac{P(A|E_i)P(E_i)}{\sum_{j=1}^{n} P(A|E_j)P(E_j)} = \frac{P(A|E_i)P(E_i)}{P(A)} \quad (15)$$

- **posterior probability:** $P(E_i|A)$
- **likelihood:** $P(A|E_i)$
- **prior:** $P(E_i)$
- **evidence (total probability):** $P(A)$

If the event A can be conditioned on only two events E_1 and E_2, then:

$$P(E_1|A) = \frac{P(A|E_1)P(E_1)}{P(A|E_1)P(E_1) + P(A|E_2)P(E_2)} \quad (16)$$

$$P(E_2|A) = \frac{P(A|E_2)P(E_2)}{P(A|E_1)P(E_1) + P(A|E_2)P(E_2)} \quad (17)$$

Example 1: Construction supplies

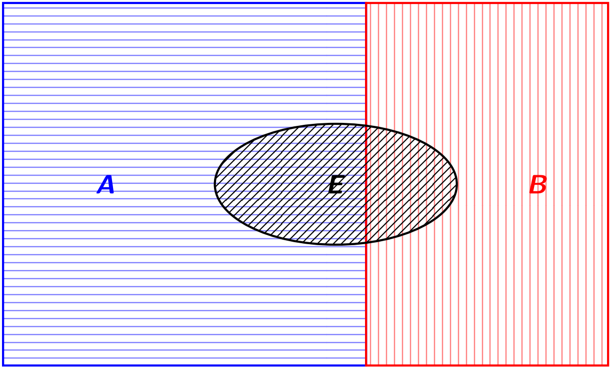
Aggregates for the construction of a reinforced concrete building are supplied by two companies. Company a delivers 600 truckloads a day while Company b delivers 400 truckloads a day. From prior experience, 3% of Company a's material is expected to be substandard while 1% of Company b's material is expected to be susbstandard.
We define:

$$\begin{aligned} A &= \text{aggregates supplied by Company } a \\ B &= \text{aggregates supplied by Company } b \\ E &= \text{aggregates are substandard} \end{aligned}$$

a. Draw a Venn diagram and convince yourself that $P(A) = 0.60, P(B) = 0.40, P(E|A) = 0.03, P(E|B) = 0.01$
b. Find the probability $P(A|E) = 0.82$.

Example 1: Construction supplies (cont.)

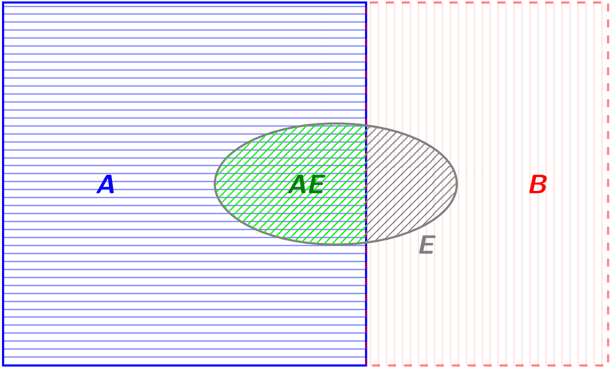
a. Draw a Venn diagram and convince yourself that $P(A) = 0.60, P(B) = 0.40, P(E|A) = 0.03, P(E|B) = 0.01$



Example 1: Construction supplies (cont.)

$$P(A) = 0.60, P(B) = 0.40, P(E|A) = 0.03, P(E|B) = 0.01$$

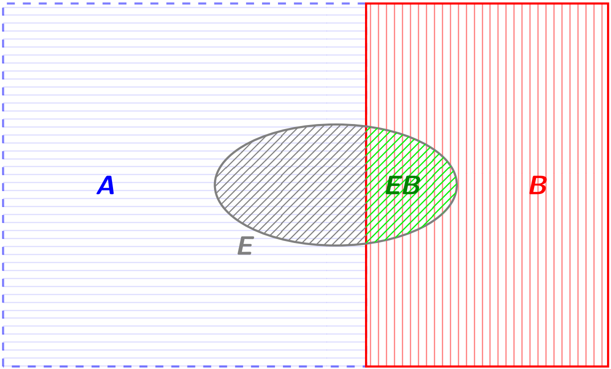
a. $P(E|A) = \frac{P(EA)}{P(A)}$



Example 1: Construction supplies (cont.)

$$P(A) = 0.60, P(B) = 0.40, P(E|A) = 0.03, P(E|B) = 0.01$$

a. $P(E|B) = \frac{P(EB)}{P(B)}$



Example 1: Construction supplies (cont.)

b) Find the probability $P(A|E) = 0.82$.

First, we find the evidence:

$$\begin{aligned} P(E) &= P(E|A)P(A) + P(E|B)P(B) \\ &= (0.03)(0.6) + (0.01)(0.4) \\ &= 0.018 + 0.004 = 0.022 \end{aligned}$$

Then we use Bayes':

$$\begin{aligned} P(A|E) &= \frac{P(E|A)P(A)}{P(E|A)P(A) + P(E|B)P(B)} \\ &= \frac{P(E|A)P(A)}{P(E)} \quad \text{(Denominator: total probability)} \\ &= \frac{0.03 \times 0.60}{0.022} \equiv \frac{\text{likelihood} \times \text{prior}}{\text{evidence}} \\ &= 0.818 \approx \boxed{0.82} \end{aligned}$$

Random variables

A random variable is a function that uniquely maps events in a sample space to the set of real numbers.

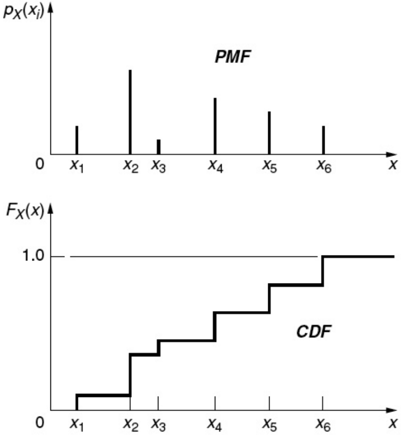
A random variable X may be:

- *Discrete*
- *Continuous*
- *Mixed* (probability defined over both discrete and range of continuous values)

Probability mass function (PMF)

The PMF is given by

$$p_X(x_i) \equiv P(X = x_i) \quad \forall x \qquad (18)$$



CDF of discrete random variable

$$F_X(x) = \sum_{x_i \leq x} P(X = x_i)$$
$$= \sum_{x_i \leq x} p_X(x_i)$$

The probability masses in a PMF sum up to 1.

Probability density function (PDF)

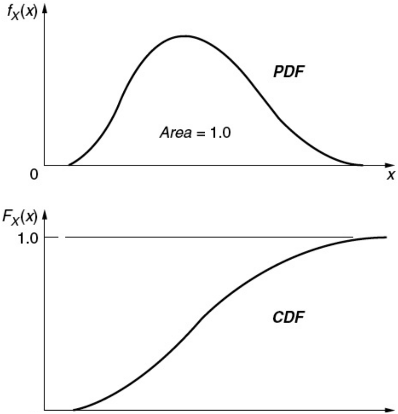
The PDF is denoted $f_X(x)$ such that the probability of X in the interval $(a, b]$ is:

$$P(a < X \leq b) = \int_a^b f_X(x)dx \quad (19)$$

CDF of continuous random variable

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(\tau)d\tau$$

It follows that the PDF is the derivative of the CDF:



The total area under a PDF is 1.

Central values

These include the mean, median and mode.

- Mean: weighted average (by probability of occurence) or expected value

$$\mathbb{E}(X) = \mu_X = \sum_i x_i p_X(x_i) \quad \text{discrete case} \tag{21}$$

$$\mathbb{E}(X) = \mu_X = \int_{-\infty}^{\infty} x f_X(x) dx \quad \text{continuous case} \tag{22}$$

Generalized expectation

The mathematical expectation can be defined for a function g of random variable X:

$$\mathbb{E}[g(X)] = \sum_i g(x_i) p_X(x_i) \quad \text{discrete case} \tag{23}$$

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) dx \quad \text{continuous case} \tag{24}$$

Measures of dispersion

Variance

In discrete case:

$$\mathbb{V}(X) = \sum_i (x_i - \mu_X)^2 p_X(x_i) \tag{25}$$

In continuous case:

$$\mathbb{V}(X) = \int_{-\infty}^{\infty} (x - \mu_X)^2 f_X(x) dx \tag{26}$$

Expanding both equations results in:

$$\mathbb{V}(X) = \mathbb{E}(X^2) - \mu_X^2 = \mathbb{E}(X^2) - \mathbb{E}(X)^2 \tag{27}$$

Measures of dispersion (cont.)

Standard deviation

The standard deviation is convenient as it has the same unit as the random variable:

$$\sigma_X = \sqrt{\mathbb{V}(X)} \tag{28}$$

Coefficient of variation

The COV gives the deviation relative to the mean. It is unitless.

$$\delta_X = \frac{\sigma_X}{\mu_X} \tag{29}$$

Mean of a linear function

For a continuous random variable X, the mean is defined as

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x f_X(x) dx \tag{30}$$

Now, given that $Z = aX + bY$, then the mean of Z is

$$\begin{aligned}
\mathbb{E}(Z) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (aX + bY) f_{X,Y} dxdy \\
&= a\int_{-\infty}^{\infty} x f_X(x) dx + b\int_{-\infty}^{\infty} y f_Y(y) dy \\
&= a\mathbb{E}(X) + b\mathbb{E}(Y)
\end{aligned}$$

Variance of a linear function

We also recall the variance of an r.v. X:

$$\mathbb{V}(X) = \mathbb{E}[(X - \mu_X)^2] \tag{31}$$

Thus, for $Z = aX + bY$:

$$\begin{aligned}
\mathbb{V}(Z) &= \mathbb{E}[((aX + bY) - (a\mu_X + b\mu_Y))^2] \\
&= \mathbb{E}[(a(X - \mu_X) + b(Y - \mu_Y))^2] \\
&= \mathbb{E}[a^2(X - \mu_X)^2 + 2ab(X - \mu_X)(Y - \mu_Y) + b^2(Y - \mu_Y)] \\
&= a^2\mathbb{V}(X) + b^2\mathbb{V}(Y) + 2abCov(X, Y)
\end{aligned}$$

Moments

The *m*-th order moment of a distribution is given by:

$$\mathbb{E}(X^m) = \begin{cases} \sum_i x_i^m \cdot p_X(x_i) & \text{(discrete)} \\ \int x^m \cdot f_X(x) dx & \text{(continuous)} \end{cases} \tag{32}$$

- *m*-th central moment: $\mathbb{E}[(X - \mu_X)^m]$
- Normalized *m*-th central moment: $\left(\frac{\mathbb{E}[(X-\mu_X)^m]}{\sigma^m}\right)$

Examples

- **Mean**: first moment, $\mathbb{E}(X)$
- **Variance**: second central moment, $\mathbb{E}[(X - \mu_X)^2]$
- **Skewness**: normalized third central moment, $\left(\frac{\mathbb{E}[(X-\mu_X)^3]}{\sigma^3}\right)$

Covariance and correlation

Recall that the variance of an r.v. X is given by:

$$\mathbb{V}(X) = \mathbb{E}[(X - \mu_X)^2] = \mathbb{E}(X^2) - \mathbb{E}(X)^2 \tag{33}$$

Then given two r.v.'s X and Y, the *covariance* measures the strength of the linear relationship between them.

Covariance

$$Cov(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)] = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) \tag{34}$$

Correlation coefficient

This is the normalized covariance

$$\rho = \frac{Cov(X, Y)}{\sigma_X \sigma_Y} \tag{35}$$

Joint distributions

Given two random variables X and Y:

Discrete case

The joint PMF is:

$$p_{X,Y}(x_i, y_j) = P(X = x_i, Y = y_j) \quad (36)$$

The CDF is:

$$F_{X,Y}(x, y) = \sum_{x_i \leq x} \sum_{y_j \leq y} p_{X,Y}(x_i, y_j) \quad (37)$$

Continuous case

The joint probability is given by:

$$P(a < X \leq b, c < Y \leq d) = \int_a^b \int_c^d f_{X,Y}(x, y) dydx \quad (38)$$

Conditional distributions of continuous random variables

Recall the definition of conditional probability (multiplication rule):

$$P(A|B) = \frac{P(AB)}{P(B)} \tag{39}$$

$$P(AB) = P(A|B)P(B) = P(B|A)P(A) \tag{40}$$

Similarly, for two continuous r.v.'s, the conditional PDF of X given Y is:

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} \tag{41}$$

Joint PDF and CDF of two variables

The joint PDF is given by:

$$f_{X,Y}(x,y) = f_{X|Y}(x|y)f_Y(y) = f_{Y|X}(y|x)f_X(x) \tag{42}$$

While the joint CDF is given by:

$$F_{X,Y}(a,b) = P(X \leq a, Y \leq b) = \int_{-\infty}^{a}\int_{-\infty}^{b} f_{X,Y}(x,y)dydx \tag{43}$$

Marginal distributions of continuous random variables

Recall the theorem of total probability:

$$P(A) = \sum_{i=1}^{n} P(A|E_i)P(E_i) \tag{44}$$

Similarly, the marginal PDFs from a joint distribution of two continuous r.v.'s X and Y is given as:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X|Y}(x|y)f_Y(y)dy = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy \tag{45}$$

$$f_Y(y) = \int_{-\infty}^{\infty} f_{Y|X}(y|x)f_X(x)dx = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dx \tag{46}$$

Bernoulli distribution

Let X be an event with only two outcomes $\{1,0\}$. And let the probability of the event be given by:

$$p(X) = \theta, \quad 0 \leq \theta \leq 1$$

And $p(X = 1) = \theta$ and $p(X = 0) = 1 - \theta$. X is said to be Bernoulli distributed:

$$X \sim \mathrm{Ber}(\theta) \tag{47}$$

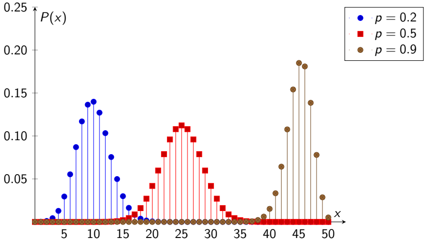
The PMF is then given by:

$$\mathrm{Ber}(x|\theta) := \theta^x (1 - \theta)^{1-x} \tag{48}$$

Binomial distribution

Given a Bernoulli sequence with X random number of occurrences of an event, N trials and θ the probability of occurrence of each event:

- $X \sim \mathrm{Bin}(N, \theta)$
- PMF: $P(X = x) := \mathrm{Bin}(x|N, \theta) := \binom{N}{x} p^x (1 - \theta)^{N-x}, \quad x = 0, 1, 2, \ldots, N$
- CDF: $F_X(x) = P(X \leq x) = \sum_{k=0}^{x} \binom{N}{k} \theta^k (1 - \theta)^{N-k}$
- Mean: $\mathbb{E}(X) = N\theta$
- Variance: $\mathbb{V}(X) = N\theta(1 - \theta)$



Bernoulli, binomial, categorical and multinomial

- The Bernoulli distribution is a special case of the binomial distribution with $N = 1$
- The categorical distribution is generalization of the Bernoulli to more than two outcomes for a single trial (e.g. set of labels $x \in \{1, ..., C\}, C > 2$):

$$\mathrm{Cat}(\boldsymbol{x}|\boldsymbol{\theta}) := \prod_{c=1}^{C} \theta_c^{x_c} \tag{49}$$

  where $\boldsymbol{x}$ is a one-hot vector (e.g. (1,0,0,0) for class 1 of four classes)
- The multinomial distribution generalizes the categorical distribution for multiple trials:

$$\mathscr{M}(\boldsymbol{x}|N, \boldsymbol{\theta}) := \binom{N}{N_1 \dots N_C} \prod_{c=1}^{C} \theta_c^{N_c} \tag{50}$$

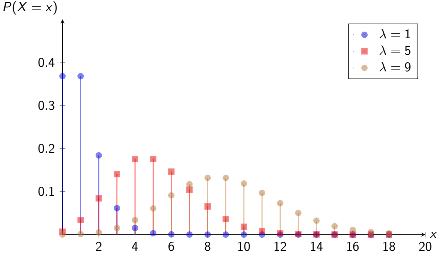
Poisson distribution

- The Poisson distribution is used to model the probability that a number of independent events occur within a fixed time interval (or within a finite space)
- Such events are described as Poisson processes
- The PMF of a Poisson random variable with **rate parameter λ** is given by:

$$P(X = x) := \mathrm{Poiss}(x|\lambda) := \frac{\lambda^x}{x!} e^{-\lambda}, \quad x \geq 0 \tag{51}$$

- The mean and variance of a Poisson random variable are equal:

$$\mathbb{E}(X) = \mathbb{V}(X) = \lambda \tag{52}$$



Gaussian distribution

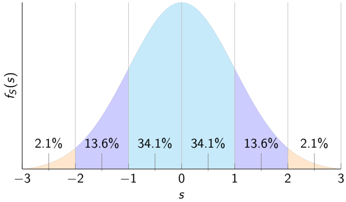
The PDF of a Gaussian (normal) distribution $X \sim \mathcal{N}(\mu, sigma^2)$ is given by:

$$\mathcal{N}(x|\mu, \sigma^2) := \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \tag{53}$$

where μ is the mean and σ^2 is the variance.

$$P(a < X \leq b) = \frac{1}{\sigma\sqrt{2\pi}} \int_a^b e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right) \tag{54}$$

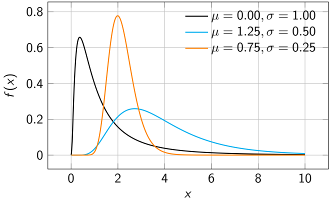
where Φ is the CDF of the standard normal distribution ($N(0,1)$).



Lognormal distribution

A random variable X that is lognormally distributed with the parameters μ and σ^2 (denoted $X \sim \mathscr{L}\mathscr{N}(\mu, \sigma^2)$) has the PDF:

$$\mathscr{L}\mathscr{N}(x|\mu, \sigma^2) = \frac{1}{(\sigma x)\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{\ln(x) - \mu}{\sigma}\right)^2\right] \quad x \geq 0 \tag{55}$$

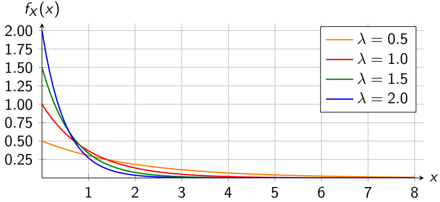


CDF: $F_X(x) = P(X \leq x) = \Phi((\ln(x) - \mu)/\sigma)$
Mean: $\mathbb{E}(X) = e^{\left(\mu + \frac{1}{2}\sigma^2\right)}$
Variance: $\mathbb{V}(X) = (e^{\sigma^2} - 1)e^{(2\mu + \sigma^2)}$

Exponential distribution

A random variable X exponentially distributed with parameter λ has the PDF:

$$\mathrm{Exp}(x|\lambda) = \lambda e^{-\lambda x} \qquad x > 0 \tag{56}$$



CDF:

$$F_X(x) = P(X \leq x) = 1 - e^{-\lambda x}, \qquad x > 0 \tag{57}$$

Mean:

$$\mathbb{E}(X) = 1/\lambda \tag{58}$$

Variance:

$$\mathbb{V}(X) = 1/\lambda^2 \tag{59}$$

Multivariate normal distribution (MVN)

The MVN PDF is given by:

$$\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) := \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right] \tag{60}$$

where:

- $\boldsymbol{\mu} = \mathbb{E}[\boldsymbol{x}] \in \mathbb{R}^D$ is the mean vector
- $\boldsymbol{\Sigma} = \mathrm{Cov}[\boldsymbol{x}]$ is the $D \times D$ covariance matrix:

$$\mathrm{Cov}[\boldsymbol{x}] := \mathbb{E}\left[(\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}])(\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}])^T\right] \tag{61}$$

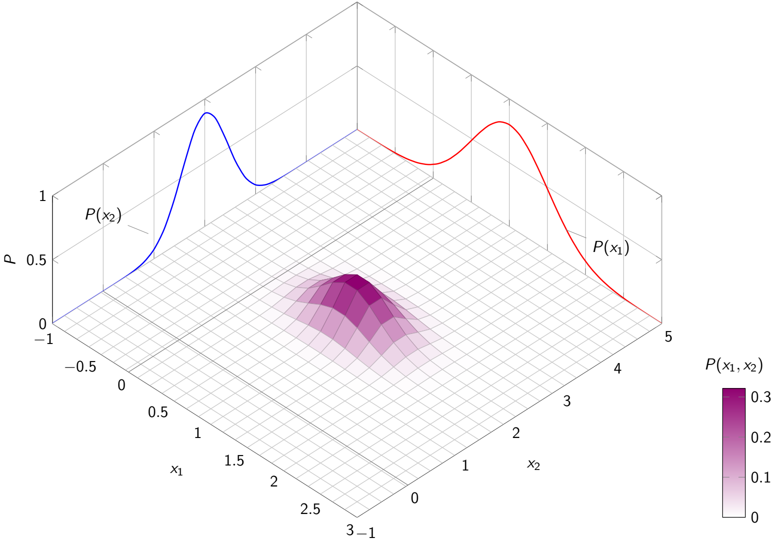
In 2D:

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12}^2 \\ \sigma_{21}^2 & \sigma_2^2 \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \tag{62}$$

where ρ is the correlation coefficient.

Bivariate MVN

Marginal distributions: $P(x_1)$ and $P(x_2)$; Joint distribution: $P(x_1, x_2)$.



Reading

- PMLI 1, 2, 3
- PMLCE 1, 3, 4