DS 5220

Supervised Machine Learning and Learning Theory

Alina Oprea
Associate Professor, CCIS
Northeastern University

September 9 2019

Class Outline

- Introduction
  – Probability and linear algebra review
- Regression - 2 weeks
  – Linear regression, polynomial, spline regression
- Classification - 4 weeks
  – Linear classification (logistic regression, LDA)
  – Non-linear models (decision trees, SVM, Naïve Bayes)
  – Ensembles (random forest, AdaBoost)
  – Model selection, regularization, cross validation
- Neural networks and deep learning – 2 weeks
  – Back-propagation, gradient descent
  – NN architectures (feed-forward, convolutional, recurrent)
- Adversarial ML – 1 lecture
  – Security of ML at testing and training time

Resources

- Instructors
  - Alina Oprea
  - TAs: Yuxuan (Ewen) Wang; Christopher Gomes
- Schedule
  - Mon, Wed 2:50-4:30pm
  - West Village H 108
  - Office hours:
    - Alina: Wed 4:30 – 6:00 pm (ISEC 625)
    - Christopher : Monday 5:00-6:00pm (ISEC 605)
    - Ewen: Thursday 5:00-6:00pm (ISEC 605)
- Online resources
  - Slides will be posted after each lecture on public website
  - Piazza for questions and discussion
  - Gradescope for homework and project submission

Classification

- Training data
  - $x_i = [x_{i,1}, \ldots x_{i,d}]$: vector of image pixels (features)
  - Size $d = 28\text{x}28 = 784$
  - y_i: image label
- Models (hypothesis)
  - Example: Linear model (parametric model)
    - $f(x) = wx + b$
  - Classify 1 if $f(x) > \mathrm{T}$; 0 otherwise
- Classification algorithm
  - Training: Learn model parameters w, b to minimize error (number of training examples for which model gives wrong label)
  - Output: "optimal" model
- Testing
  - Apply learned model to new data and generate prediction $f(x)$

Supervised Learning: Classification

Training

Data → Pre-processing → Feature extraction → Learning model

Labeled

$x_i, y_i \in \{0,1\}$

Normalization

Feature Selection

Classification

$f(x)$

Testing

New data → Learning model → Predictions $y' = f(x') \in \{0,1\}$

Unlabeled

x'

$f(x)$

Positive
Negative

Classification

Supervised Learning: Regression

Training

Data → Pre-processing → Feature extraction → Learning model

Labeled

$x_i, y_i \in R$

Normalization

Feature Selection

Regression

$f(x)$

Testing

New data → Learning model → Predictions

$y' = f(x') \in R$

Unlabeled

x'

$f(x)$

Response variable

Regression

Supervised Learning Tasks

- Classification
  - Learn to predict class (discrete)
  - Minimize error 1/N $\sum_{i=1}^{N}[y_i \neq f(x_i)]$
- Regression
  - Learn to predict response variable (numerical)
  - Minimize MSE (Mean Square Error between prediction and actual values): loss function
- Both classification and regression
  - Training and testing phase
  - “Optimal” model is learned in training and applied in testing

Terminology

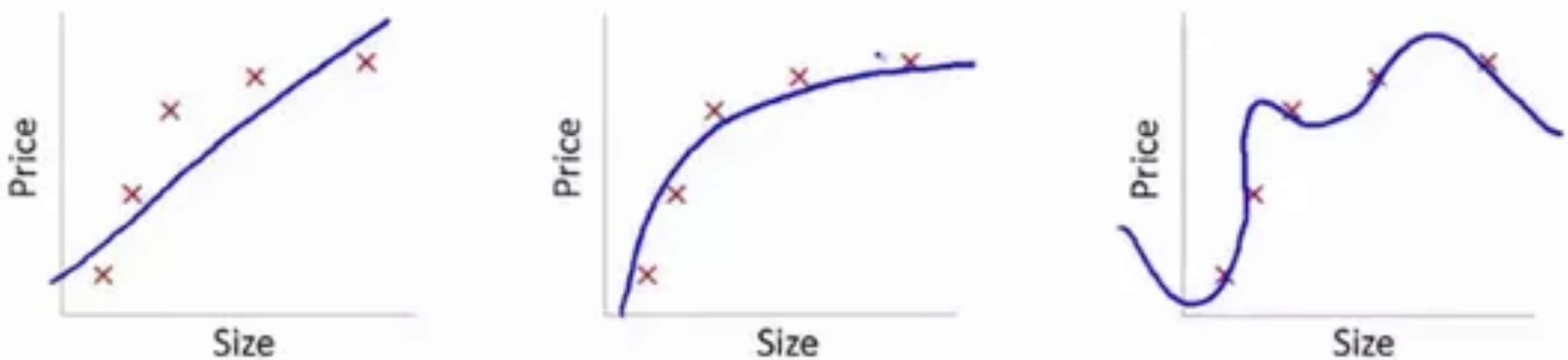
- Hypothesis space $H = \{f: X \to Y\}$
- Training data $\mathrm{D} = (x_\mathrm{i}, y_\mathrm{i}) \in X \times Y$
- Features: $x_i \in X$
- Labels / response variables $y_\mathrm{i} \in Y$
  - Classification: discrete $y_\mathrm{i} \in \{0,1\}$
  - Regression: $y_\mathrm{i} \in \mathrm{R}$
- Loss function: $L(f, D)$
  - Measures how well f fits training data
- Training algorithm: Find hypothesis $\hat{f}: X \to Y$
  - $\hat{f} = \underset{f \in H}{\operatorname{argmin}}\ L(f, D)$

Learning Challenges

- Goal
  - Classify well new testing data
  - Model generalizes well to new testing data
- Variance
  - Amount by which model would change if we estimated it using a different training data set
  - More complex models result in higher variance
- Bias
  - Error introduced by approximating a real-life problem by a much simpler model
  - E.g., assume linear model (linear regression)
  - More complex models result in lower bias

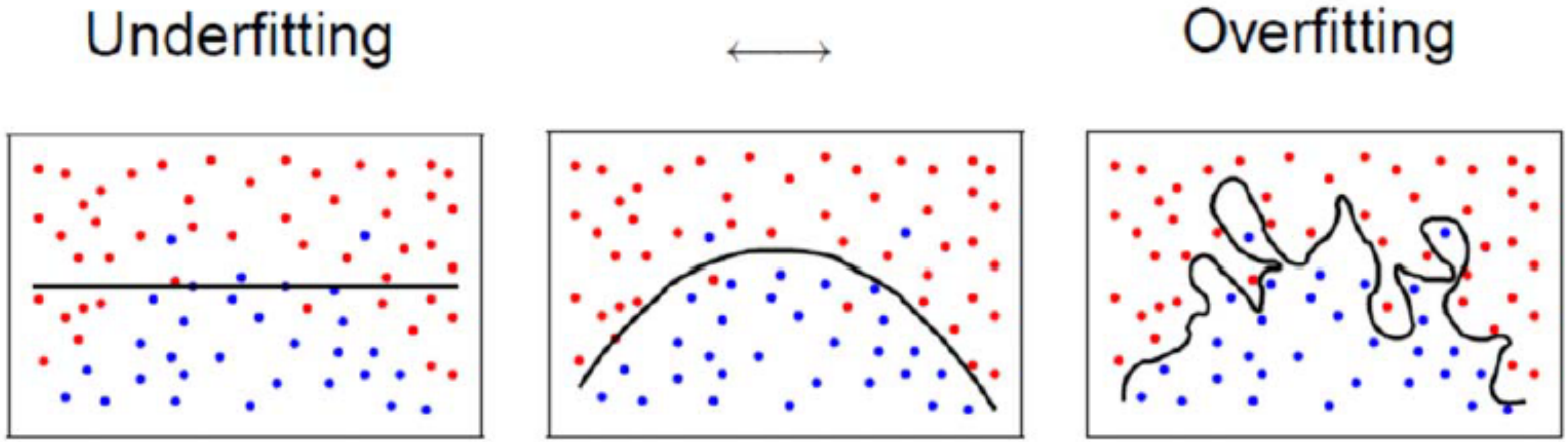
Bias-Variance tradeoff

Example: Regression



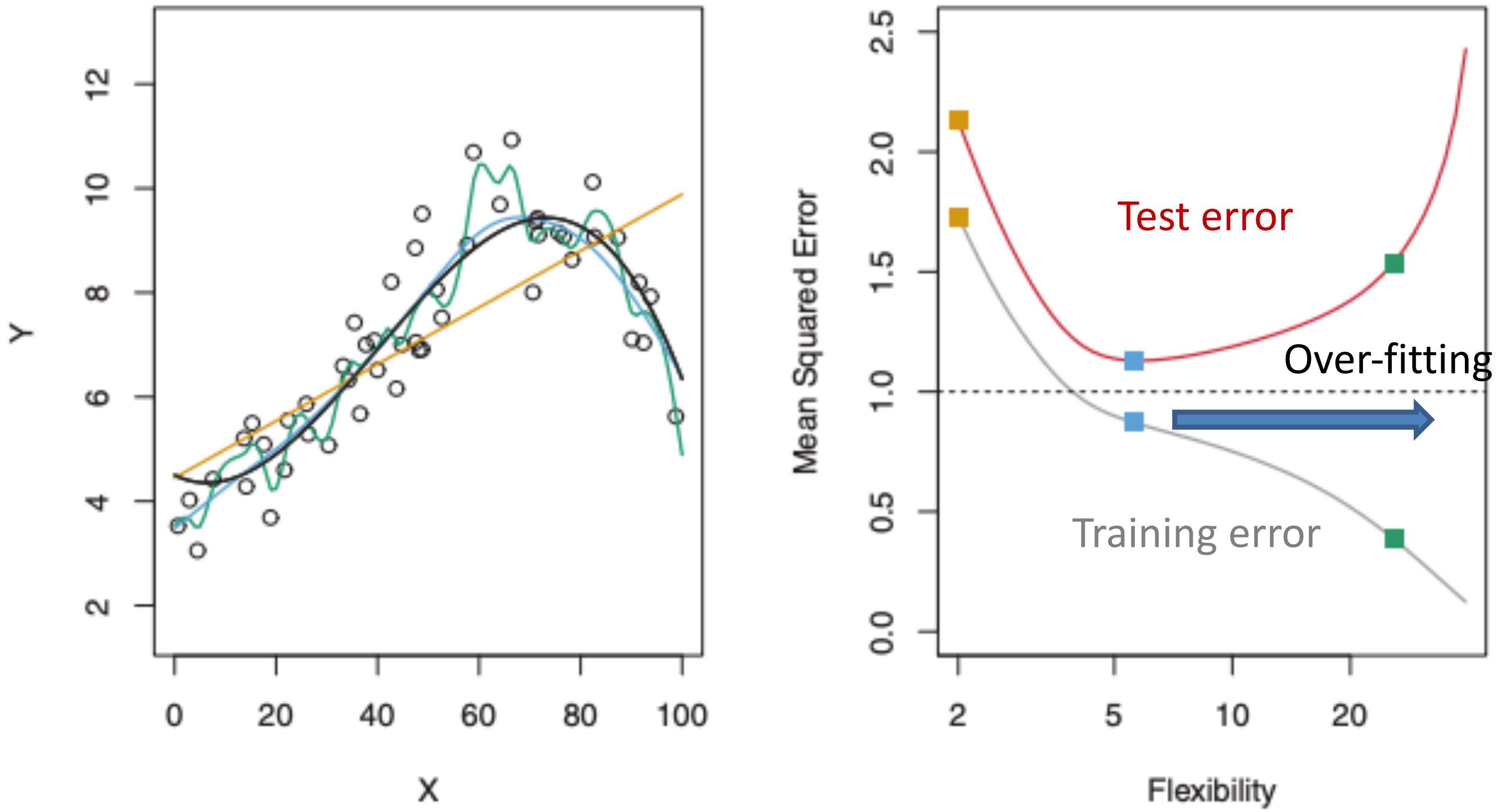
$\theta_0 + \theta_1 x$	$\theta_0 + \theta_1 x + \theta_2 x^2$	$\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$
High bias (underfit)	"Just right"	High variance (overfit)

Generalization Problem in Classification

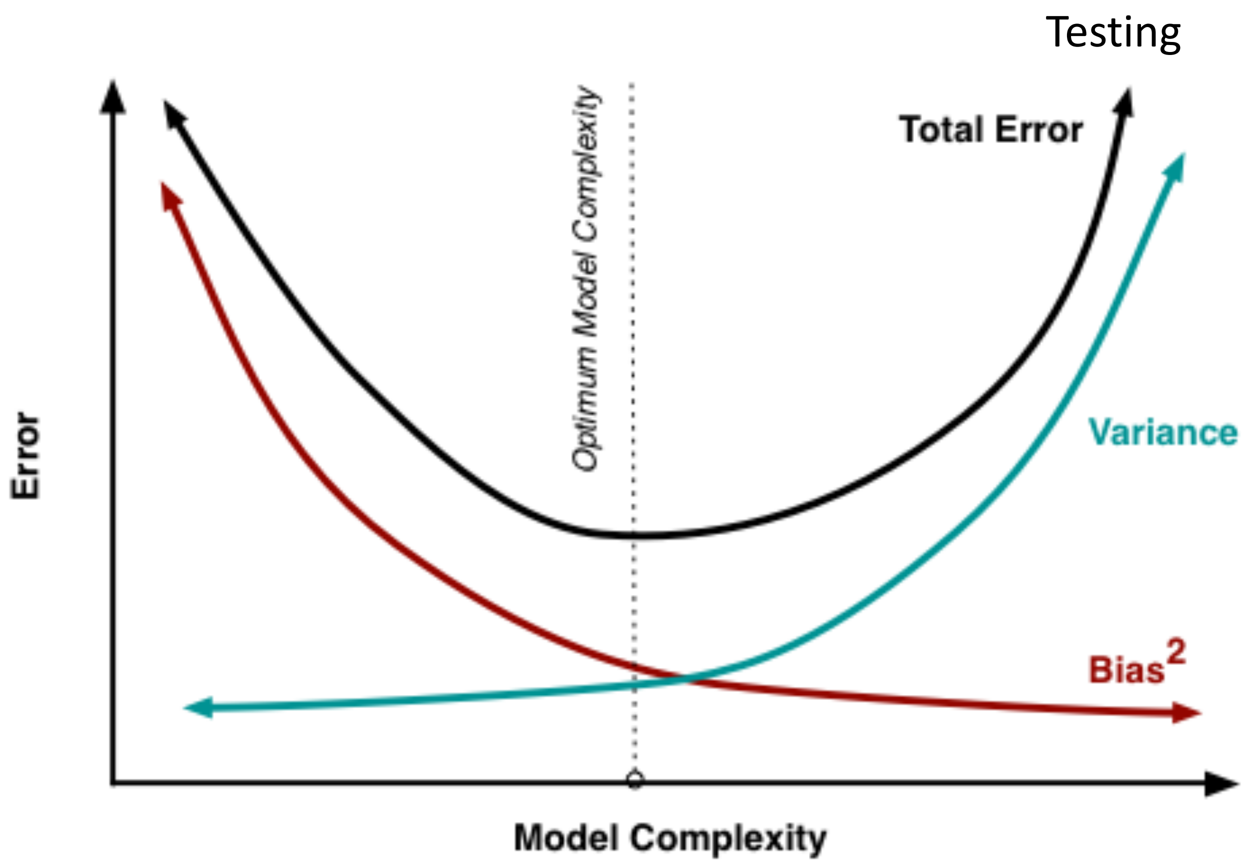


• Again, need to control the complexity of the (discriminant) function

Training and testing error



Bias-Variance Tradeoff



Model underfits the data

Model overfits the data

Occam's Razor

- William of Occam: Monk living in the 14^{th} century
- Principle of parsimony:

"One should not increase, beyond what is necessary, the number of entities required to explain anything"

- When many solutions are available for a given problem, we should select the simplest one
- But what do we mean by simple?
- We will use prior knowledge of the problem to solve to define what is a simple solution

Summary

- ML is a subset of AI designing learning algorithms
- Learning tasks are *supervised* (e.g., classification and regression) or *unsupervised* (e.g., clustering)
  - Supervised learning uses labeled training data
- Learning the “best” model is challenging
  - Design algorithm to fit the data
  - Bias-Variance tradeoff
  - Need to generalize on new, unseen test data
  - Occam’s razor (prefer simplest model with good performance)

Probability review

Probability Resources

- Review notes from Stanford's machine learning class
- Sam Roweis's probability review
- David Blei’s probability review
- Books:
  – Sheldon Ross, A First course in probability

Discrete Random Variables

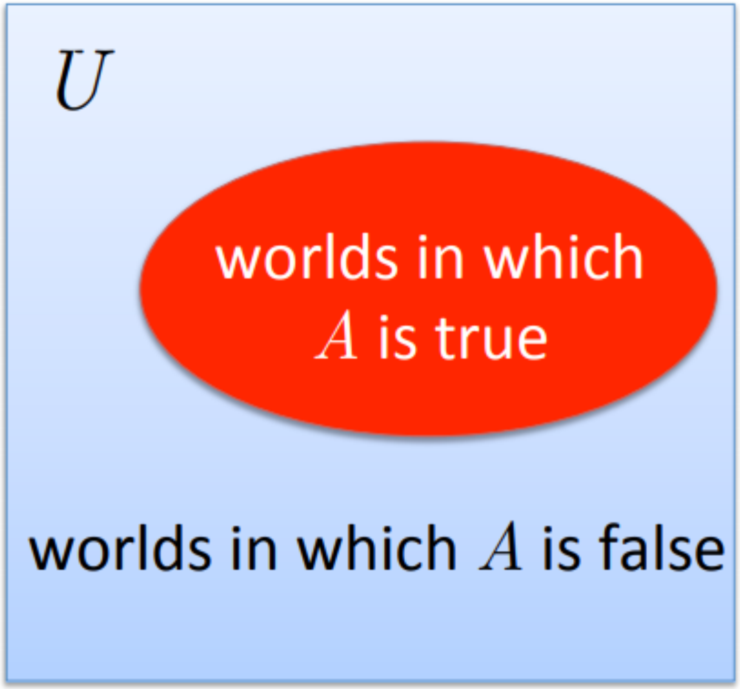
- Let A denote a random variable
  - A represents an event that can take on certain values
  - Each value has an associated probability
- Examples of binary random variables:
  - A = I have a headache
  - A = Sally will be the US president in 2020
- $\mathrm{P}(A)$ is "the fraction of possible worlds in which A is true"

Visualizing A

- Universe U is the event space of all possible worlds
  - Its area is 1
  - $\mathrm{P}(U) = 1$
- $\mathrm{P}(A)$ = area of red oval
- Therefore:

$$P(A) + P(\neg A) = 1$$

$$P(\neg A) = 1 - P(A)$$



Axioms of Probability

Kolmogorov showed that three simple axioms lead to the rules of probability theory

- de Finetti, Cox, and Carnap have also provided compelling arguments for these axioms

1. All probabilities are between 0 and 1:
   $$0 \leq \mathrm{P}(A) \leq 1$$
2. Valid propositions (tautologies) have probability 1, and unsatisfiable propositions have probability 0:
   $$\mathrm{P}(\text{true}) = 1\,; \quad \mathrm{P}(\text{false}) = 0$$
3. The probability of a disjunction is given by:
   $$\mathrm{P}(A \vee B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \wedge B)$$

Interpreting the Axioms

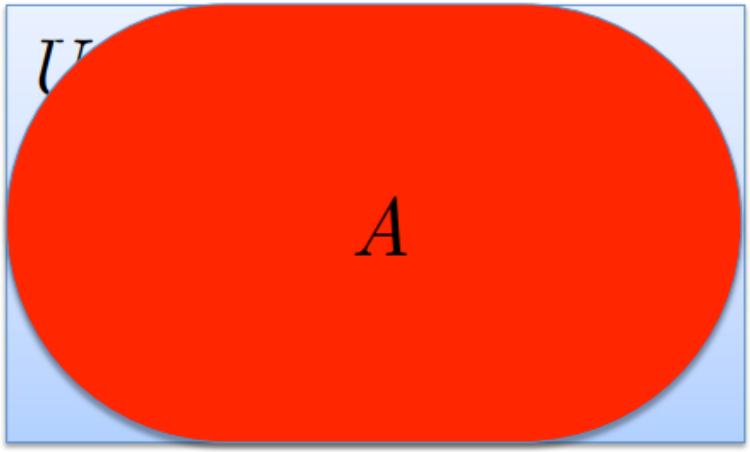
- $0 \leq \mathrm{P}(A) \leq 1$
- $\mathrm{P}(\text{true}) = 1$
- $\mathrm{P}(\text{false}) = 0$
- $\mathrm{P}(A \vee B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \wedge B)$

The area of A can't get any smaller than 0

A zero area would mean no world could ever have A true

Interpreting the Axioms

- $0 \leq P(A) \leq 1$
- $P(\text{true}) = 1$
- $P(\text{false}) = 0$
- $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$

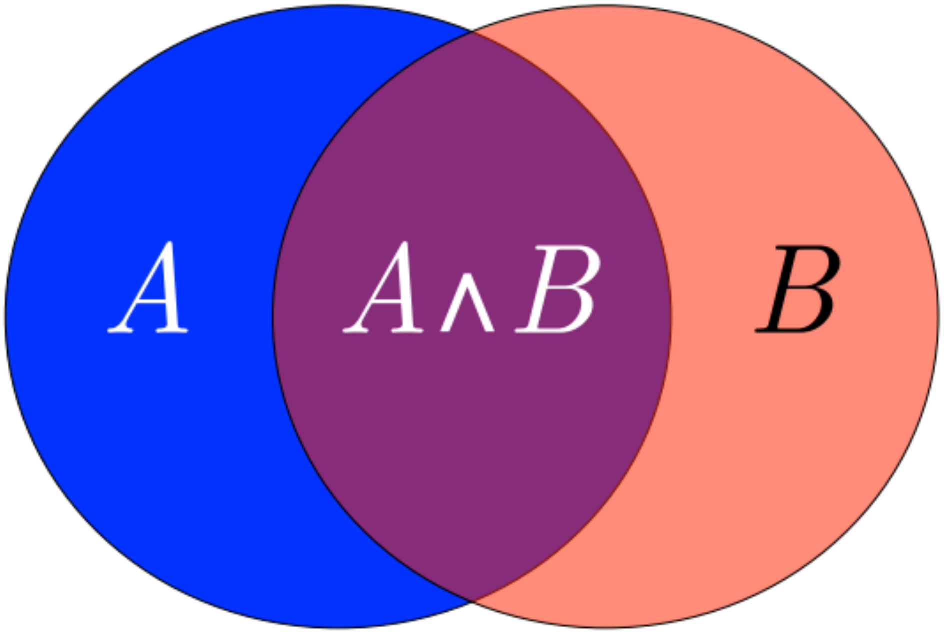


The area of A can't get any bigger than 1

An area of 1 would mean A is true in all possible worlds

Interpreting the Axioms

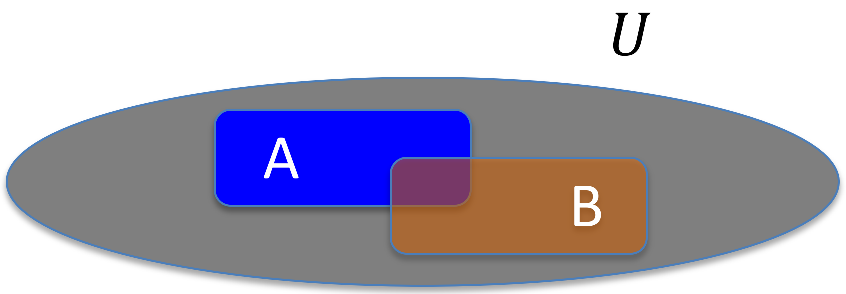
- $0 \leq \mathrm{P}(A) \leq 1$
- $\mathrm{P}(\text{true}) = 1$
- $\mathrm{P}(\text{false}) = 0$
- $\mathrm{P}(A \vee B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \wedge B)$



The union bound

- For events A and B

$$P[\, A \cup B \,] \le P[A] + P[B]$$



Axiom: $P[\, A \cup B \,] = P[A] + P[B] - P[A \cap B]$

If $A \cap B = \Phi$, then $P[\, A \cup B \,] = P[A] + P[B]$

Example:

$A_1 = \{$ all x in $\{0,1\}^n$ s.t $lsb_2(x)=11$ $\}$; $A_2 = \{$ all x in $\{0,1\}^n$ s.t. $msb_2(x)=11$ $\}$

$P[\, lsb_2(x)=11$ or $msb_2(x)=11 \,] = P[A_1 \cup A_2] \le ¼+¼ = ½$

Random Variables (Discrete)

Def: a random variable X is a function $X: U \longrightarrow V$

Def: A discrete random variable takes a finite number of values: $|V|$ is finite

Example: X is modeling a coin toss with output 1 (heads) or 0 (tail)

$P[X=1] = p$, $P[X=0] = 1-p$

Bernoulli Random Variable

We write $X \longleftarrow U$ to denote a **uniform random variable (discrete)** over U

for all $u \in U$: $P[\, X = u \,] = 1/|U|$

Example: If $p=1/2$; then X is a uniform coin toss

Probability Mass Function (PMF): $p(v) = P[X = v]$

Example

1. X is the number of heads in a sequence of n coin tosses

What is the probability $\mathrm{P}[X = k]$?

$$\mathrm{P}[X = k] = \binom{n}{k} p^k (1-p)^{n-k}$$

Binomial Random Variable

2. X is the sum of two fair dice

What is the probability $\mathrm{P}[X = k]$ for $k \in \{2, \dots, 12\}$?

P[X=2]=1/36; P[X=3]=2/36; P[X=4]= 3/36

For what k is $\mathrm{P}[X = k]$ highest?

Multi-Value Random Variable

- Suppose A can take on more than 2 values
- A is a *random variable with arity* k if it can take on exactly one value out of $\{v_1, v_2, \ldots, v_k\}$
- Thus…

$$P(A = v_i \wedge A = v_j) = 0 \quad \text{if } i \neq j$$

$$P(A = v_1 \vee A = v_2 \vee \ldots \vee A = v_k) = 1$$

$$\boxed{1 = \sum_{i=1}^{k} P(A = v_i)}$$

Multi-Value Random Variable

- We can also show that:

$$P(B) = P(B \wedge [A = v_1 \vee A = v_2 \vee \ldots \vee A = v_k])$$

$$\boxed{P(B) = \sum_{i=1}^{k} P(B \wedge A = v_i)}$$

- This is called **marginalization** over A

Expectation and variance

Expectation for discrete random variable X

$$E[X] = \sum_{v} vP[X = v]$$

Properties

- $E[aX] = a\,E[X]$
- Linearity: $E[X + Y] = E[X] + E[Y]$

Variance

$$Var[X] \triangleq E[(X - E(X))^2]$$

$$\begin{aligned} E[(X - E[X])^2] &= E[X^2 - 2E[X]X + E[X]^2] \\ &= E[X^2] - 2E[X]E[X] + E[X]^2 \\ &= E[X^2] - E[X]^2, \end{aligned}$$

Example discrete RVs

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

$$p(x) = p(1-p)^{x-1}$$

Continuous Random Variables

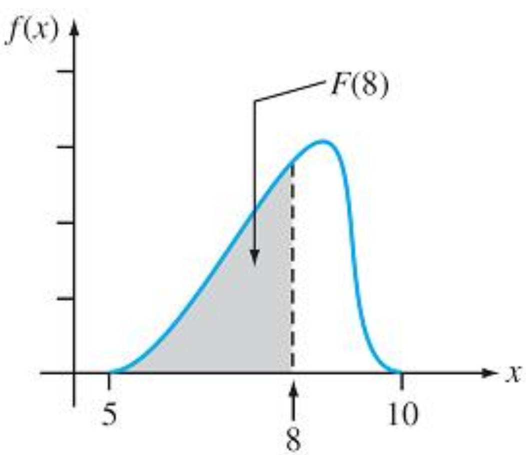
- X:U⟶V is continuous RV if it takes infinite number of values
- The **cumulative distribution function CDF** *F: R* ⟶ {0,1} for *X* is defined for every value *x* by:

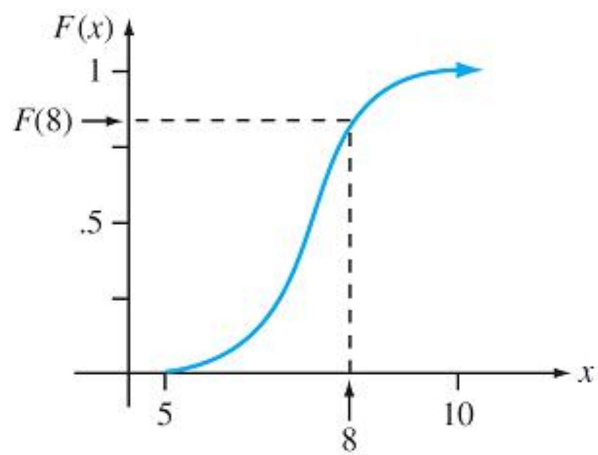
$$F(x) = \Pr(X \leq x)$$

- The **probability distribution function PDF** *f*(*x*) for *X is*

$$f(x) = dF(x)/dx$$

Increasing





Example continuous RV

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

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

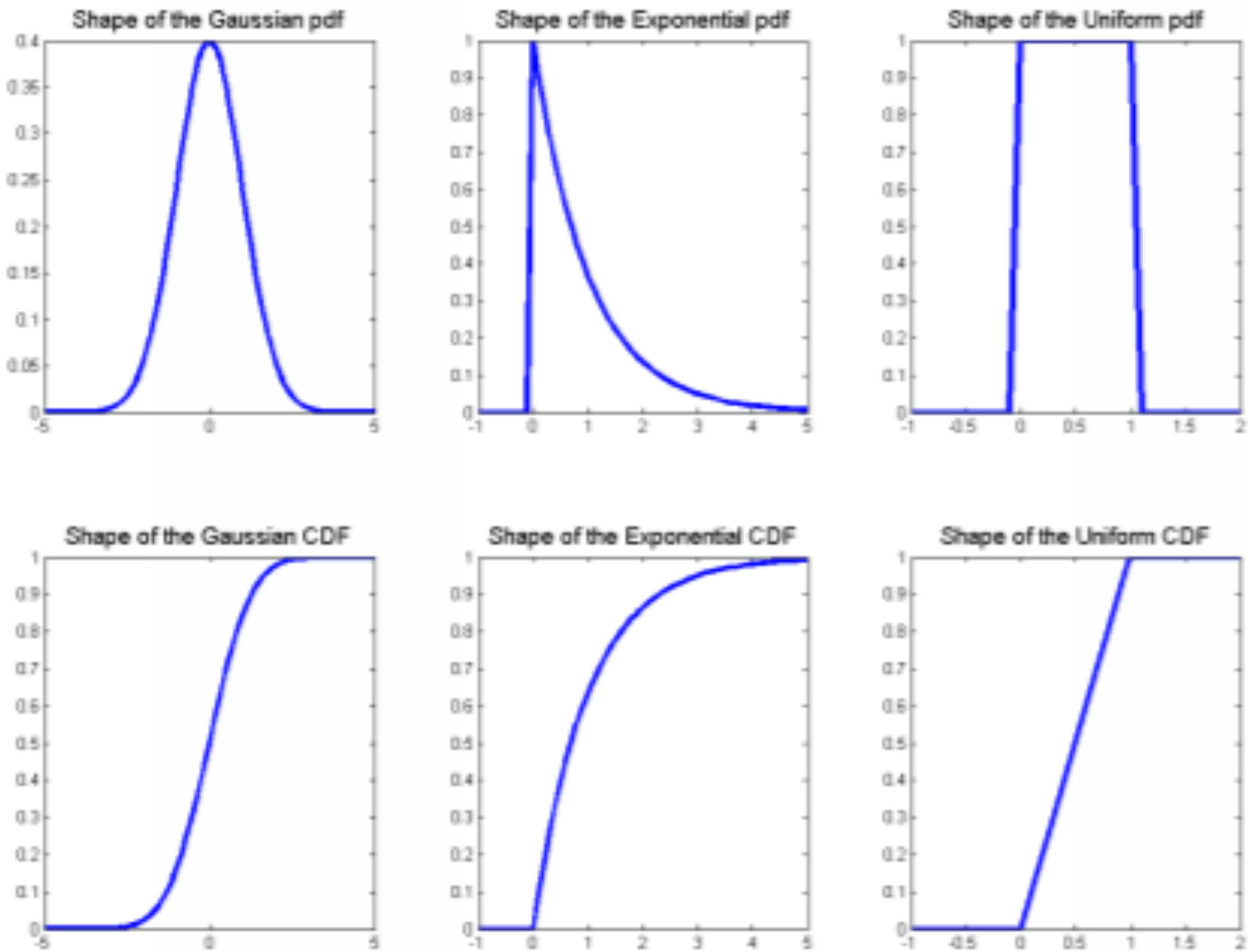
- $X \sim Exponential(\lambda)$ (where $\lambda > 0$): decaying probability density over the nonnegative reals.

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

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

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

Example CDFs and PDFs



Continuous RV

Expectation for continuous random variable X

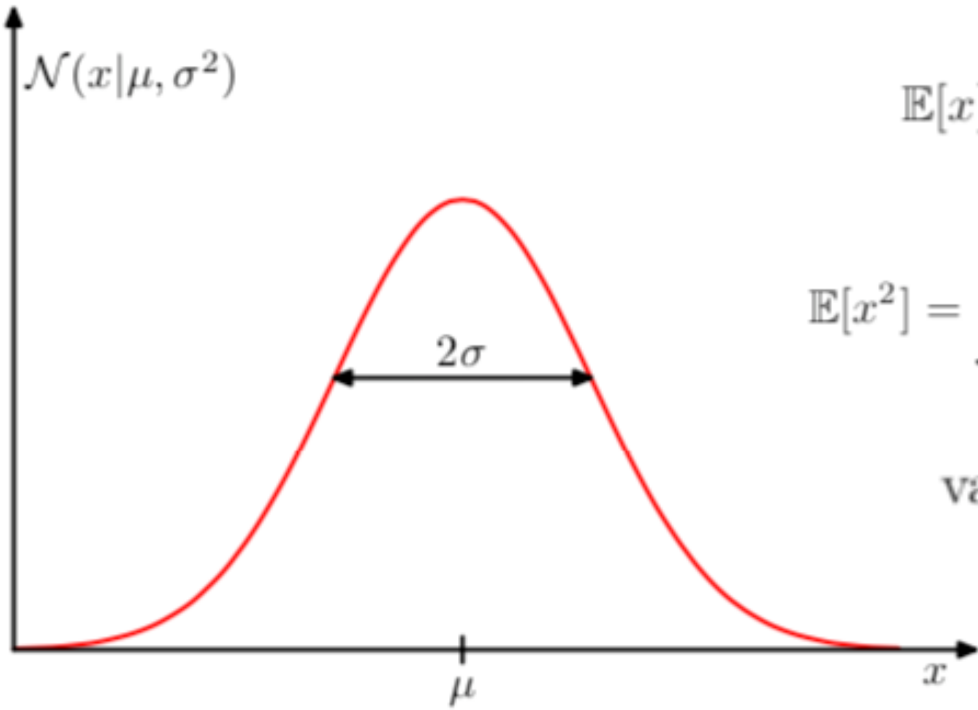
$$E[g(X)] \triangleq \int_{-\infty}^{\infty} g(x) f_X(x) dx.$$

Variance is similar!

Example: Let X be uniform RV on [a,b]

- What is the CDF and PDF?
- Compute the expectation and variance of X

Normal (Gaussian) distribution



$$\mathbb{E}[x] = \int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu,\sigma^2\right) x \, \mathrm{d}x = \mu.$$

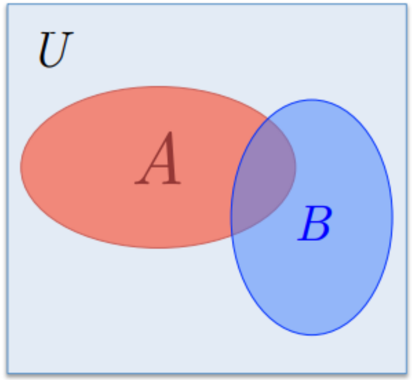
$$\mathbb{E}[x^2] = \int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu,\sigma^2\right) x^2 \, \mathrm{d}x = \mu^2 + \sigma^2.$$

$$\mathrm{var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2 = \sigma^2$$

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

Conditional Probability

- $\mathrm{P}(A \mid B)$ = Fraction of worlds in which B is true that also have A true



What if we already know that B is true?

That knowledge changes the probability of A

- Because we know we're in a world where B is true

$$P(A \mid B) = \frac{P(A \wedge B)}{P(B)}$$

$$P(A \wedge B) = P(A \mid B) \times P(B)$$

<u>Def</u>: Events A and B are **independent** if and only if

$$\mathrm{P}[\,A \cap B\,] = \mathrm{P}[A] \cdot \mathrm{P}[B]$$

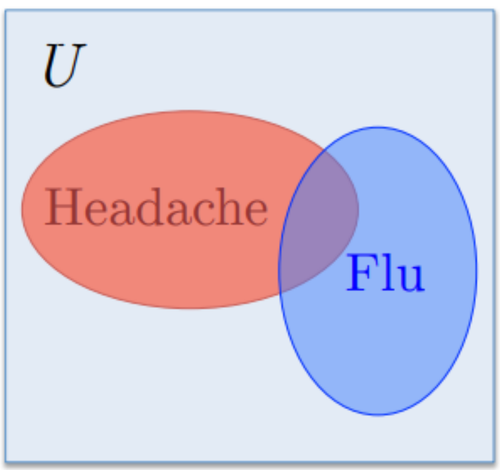
If A and B are independent

$$\mathrm{P}[A|B] = \frac{\mathrm{P}[A \cap B]}{\mathrm{P}[B]} = \frac{\mathrm{P}[A]\mathrm{P}[B]}{\mathrm{P}[B]} = \mathrm{P}[\mathrm{A}]$$

Inference from Conditional Probability

$$P(A \mid B) = \frac{P(A \wedge B)}{P(B)}$$

$$P(A \wedge B) = P(A \mid B) \times P(B)$$



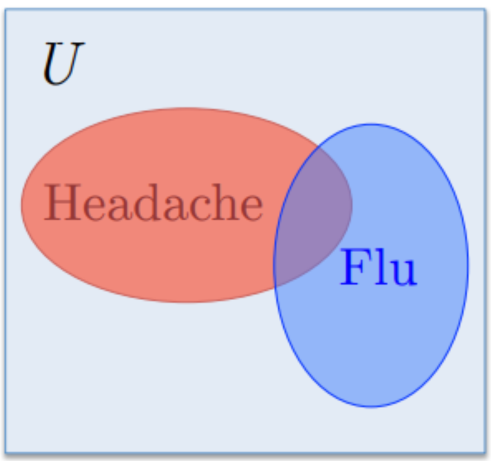
P(headache) = 1/10
P(flu) = 1/40
P(headache | flu) = 1/2

"Headaches are rare and flu is rarer, but if you're coming down with the flu there's a 50-50 chance you'll have a headache."

Inference from Conditional Probability

$$P(A \mid B) = \frac{P(A \wedge B)}{P(B)}$$

$$P(A \wedge B) = P(A \mid B) \times P(B)$$



P(headache) = 1/10
P(flu) = 1/40
P(headache | flu) = 1/2

One day you wake up with a headache. You think: "Drat! 50% of flus are associated with headaches so I must have a 50-50 chance of coming down with flu."

Is this reasoning good?

Inference from Conditional Probability

$$P(A \mid B) = \frac{P(A \wedge B)}{P(B)}$$

$$P(A \wedge B) = P(A \mid B) \times P(B)$$

$P(\text{headache}) = 1/10$
$P(\text{flu}) = 1/40$
$P(\text{headache} \mid \text{flu}) = 1/2$

Want to solve for:
$P(\text{headache} \wedge \text{flu}) = ?$
$P(\text{flu} \mid \text{headache}) = ?$

$$\begin{aligned} P(\text{headache} \wedge \text{flu}) &= P(\text{headache} \mid \text{flu}) \text{ x } P(\text{flu}) \\ &= 1/2 \text{ x } 1/40 = 0.0125 \end{aligned}$$

$$\begin{aligned} P(\text{flu} \mid \text{headache}) &= P(\text{headache} \wedge \text{flu}) \,/\, P(\text{headache}) \\ &= 0.0125 \,/\, 0.1 = 0.125 \end{aligned}$$

Bayes Theorem

Bayes' Rule

$$P(A \mid B) = \frac{P(B \mid A) \times P(A)}{P(B)}$$

- Exactly the process we just used
- The most important formula in probabilistic machine learning

Bayes, Thomas (1763) An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London,* **53:370-418**

(Super Easy) Derivation:

$$\boxed{\begin{matrix} P(A \wedge B) \\ P(B \wedge A) \end{matrix}} \begin{matrix} = P(A \mid B) \times P(B) \\ = P(B \mid A) \times P(A) \end{matrix}$$

these are the same

Just set equal...

$$P(A \mid B) \times P(B) = P(B \mid A) \times P(A)$$

and solve...

Linear algebra review

Resources

- Zico Kolter, Linear algebra review
- Sam Roweis's linear algebra review
- Books:
  - O. Bretscher, Linear Algebra with Applications

Vectors and matrices

- Vector in R^n is an ordered set of n real numbers.
  - e.g. v = (1,6,3,4) is in R^4
  - A column vector: $\begin{pmatrix} 1 \\ 6 \\ 3 \\ 4 \end{pmatrix}$
  - A row vector: $\begin{pmatrix} 1 & 6 & 3 & 4 \end{pmatrix}$
- m-by-n matrix is an object in R^{mxn} with m rows and n columns, each entry filled with a (typically) real number:

$$\begin{pmatrix} 1 & 2 & 8 \\ 4 & 78 & 6 \\ 9 & 3 & 2 \end{pmatrix}$$

Matrix multiplication

We will use upper case letters for matrices. The elements are referred by $A_{i,j}$.

- Matrix product:

$$A \in \mathbb{R}^{m \times n} \qquad B \in \mathbb{R}^{n \times p}$$

$$C = AB \in \mathbb{R}^{m \times p}$$

$$C_{ij} = \sum_{k=1}^{n} A_{ik} B_{kj}$$

e.g.

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}, B = \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix}$$

$$AB = \begin{pmatrix} a_{11}b_{11} + a_{12}b_{21} & a_{11}b_{12} + a_{12}b_{22} \\ a_{21}b_{11} + a_{22}b_{21} & a_{21}b_{12} + a_{22}b_{22} \end{pmatrix}$$

Matrix transpose

Transpose: You can think of it as

– "flipping" the rows and columns

OR

– "reflecting" vector/matrix on line

e.g. $\begin{pmatrix} a \\ b \end{pmatrix}^T = \begin{pmatrix} a & b \end{pmatrix}$

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^T = \begin{pmatrix} a & c \\ b & d \end{pmatrix}$$

- $(A^T)^T = A$
- $(AB)^T = B^T A^T$
- $(A+B)^T = A^T + B^T$

A is a symmetric matrix if $A = A^T$

Inverse of a matrix

- Inverse of a square matrix A, denoted by A^{-1} is the *unique* matrix s.t.
  – $AA^{-1} = A^{-1}A = I$ (identity matrix)

- If A^{-1} and B^{-1} exist, then
  – $(AB)^{-1} = B^{-1}A^{-1}$,
  – $(A^T)^{-1} = (A^{-1})^T$

- For diagonal matrices $\mathbf{D}^{-1} = \mathrm{diag}\{d_1^{-1}, \ldots, d_n^{-1}\}$

Linear independence

- A set of vectors is linearly independent if none of them can be written as a linear combination of the others.
- Vectors $v_1,\ldots,v_k$ are linearly independent if $c_1v_1+\ldots+c_kv_k = 0$ implies $c_1=\ldots=c_k=0$
- Otherwise they are linearly dependent

$$\begin{pmatrix} | & | & | \\ v_1 & v_2 & v_3 \\ | & | & | \end{pmatrix}\begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

e.g.

$$\begin{pmatrix} 1 & 0 \\ 2 & 3 \\ 1 & 3 \end{pmatrix}$$

(c_1, c_2)=(0,0), i.e. the columns are linearly independent.

$$x_1 = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \quad x_2 = \begin{bmatrix} 4 \\ 1 \\ 5 \end{bmatrix} \quad x_3 = \begin{bmatrix} 2 \\ -3 \\ -1 \end{bmatrix}$$

Linearly dependent

$$x_3 = -2x_1 + x_2$$

Rank of a Matrix

- rank(A) (the rank of a m-by-n matrix A) is

  The maximal number of linearly independent columns

  The maximal number of linearly independent rows

- If A is n by m, then
  - rank(A)<= min(m,n)

- Examples

$$\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \quad \begin{pmatrix} 2 & 1 \\ 4 & 2 \end{pmatrix} \quad \begin{pmatrix} 2 & 1 & 3 \\ 0 & 5 & 2 \end{pmatrix}$$

System of linear equations

$$\begin{array}{rcrcl} 4x_1 & - & 5x_2 & = & -13 \\ -2x_1 & + & 3x_2 & = & 9. \end{array}$$

Matrix formulation

$$Ax = b$$

$$A = \begin{bmatrix} 4 & -5 \\ -2 & 3 \end{bmatrix}, \quad b = \begin{bmatrix} -13 \\ 9 \end{bmatrix}.$$

If A has an inverse, solution is $x = A^{-1}b$

Covariance

- X and Y are random variables
- $Cov(X,Y) = E[(X - E(X))(Y - E(Y))]$
- Properties

$$\text{(i)}\quad \mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$$

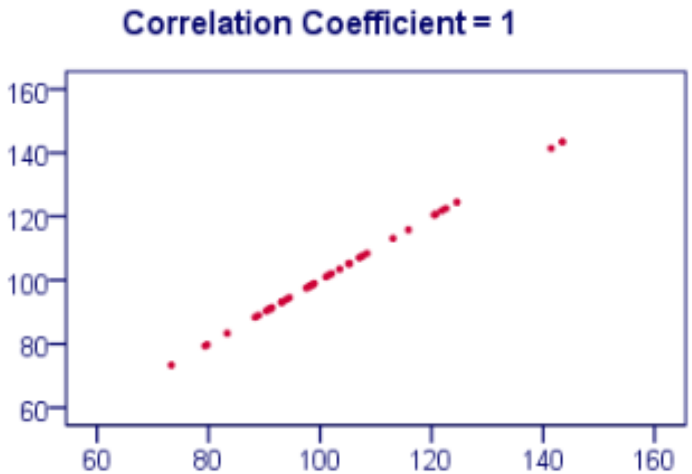
$$\text{(ii)}\quad \mathrm{Cov}(X,X) = \mathrm{Var(X)}$$

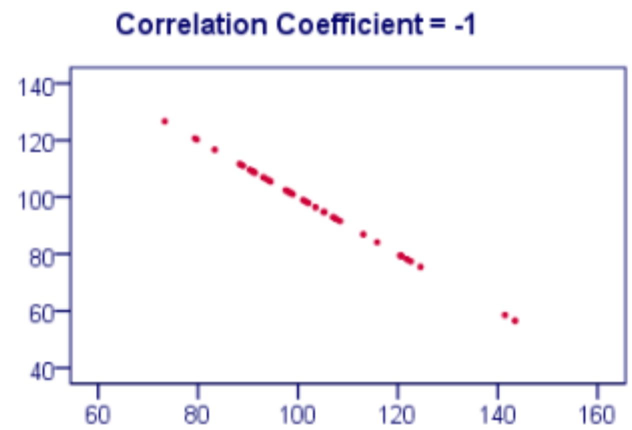
$$\text{(iii)}\quad \mathrm{Cov}(aX,Y) = a\,\mathrm{Cov}(X,Y)$$

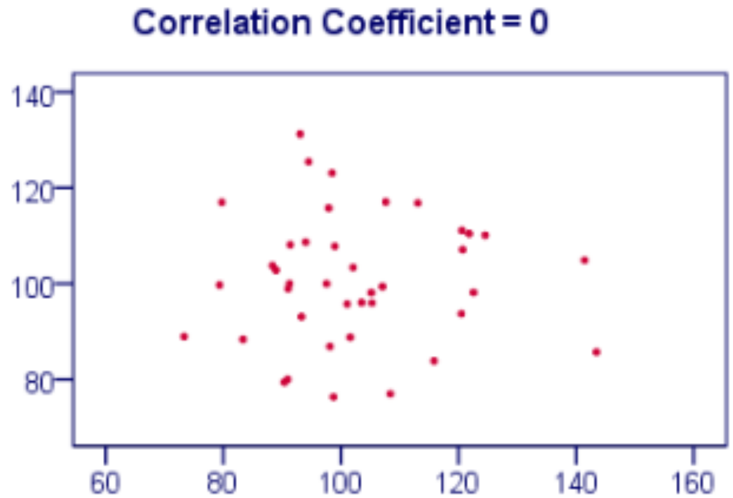
$$\text{(iv)}\quad \mathrm{Cov}\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} \mathrm{Cov}(X_i, Y_j)$$

Pearson Correlation

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

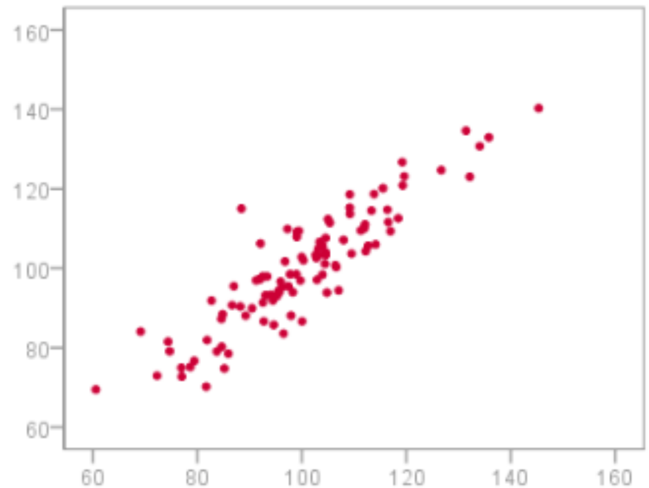




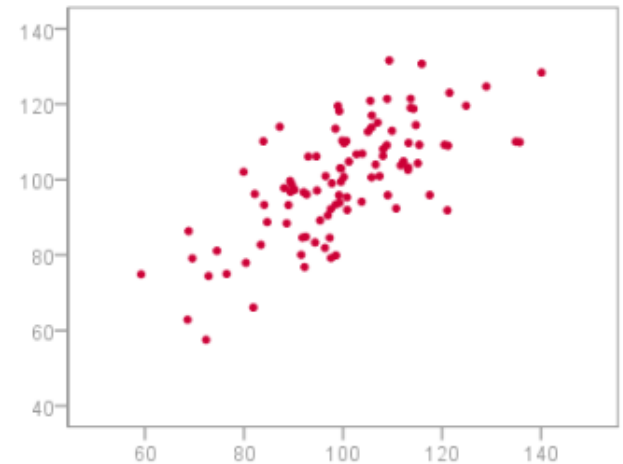


Positive/Negative Correlation

Positive Correlation

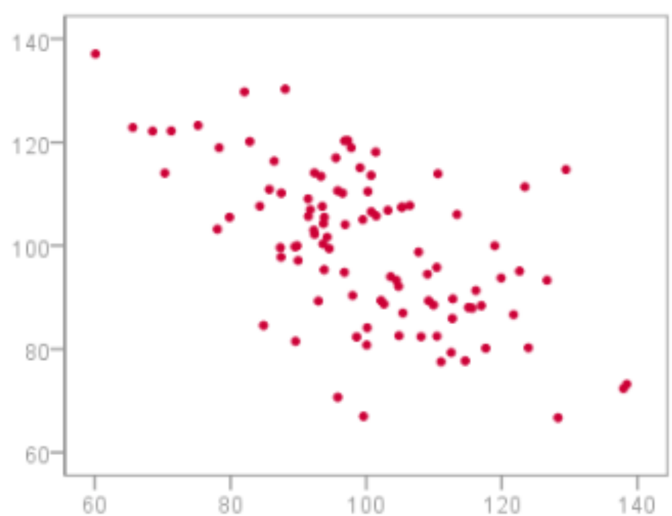


r = 0.9

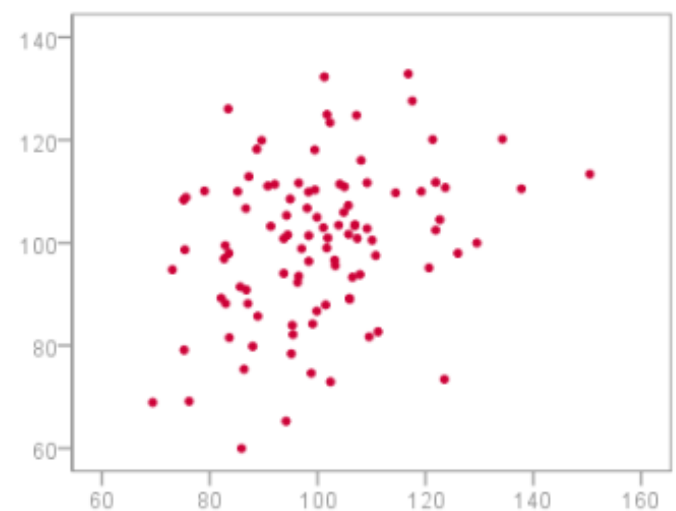


r = 0.7

Negative Correlation



r = -0.6



r = 0.3

Bivariate normal

$$F_{X,Y}(x, y) = \mathrm{P}(X \le x, Y \le y) \quad \textbf{(Eq.1)}$$

Joint CDF

- $X \sim N(\mu_X, \sigma_X)$ and $Y \sim N(\mu_Y, \sigma_Y)$ are Normal
- $\mu = (E[X], E[Y]) = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}$

  mean vector

- $\Sigma = \begin{pmatrix} Var(X) & Cov(X,Y) \\ Cov(X,Y) & Var(Y) \end{pmatrix} =$

  $\begin{pmatrix} \sigma_X^2 & \rho\, \sigma_X \sigma_Y \\ \rho\, \sigma_X \sigma_Y & \sigma_Y^2 \end{pmatrix}$

  covariance matrix

Bivariate normal

$$\boxed{F_{X,Y}(x,y) = \mathrm{P}(X \leq x, Y \leq y) \quad \textbf{(Eq.1)}}$$

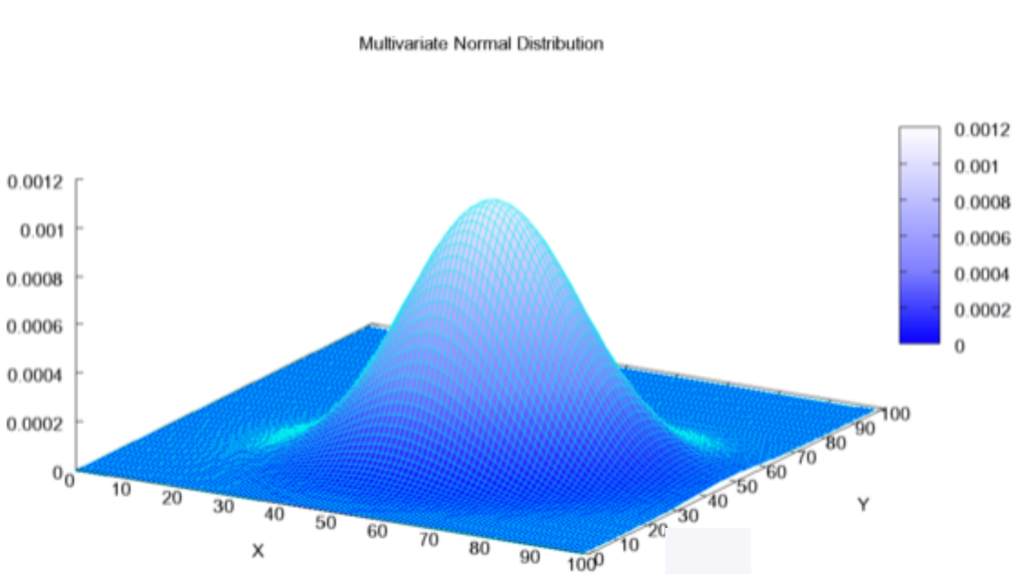
Joint CDF

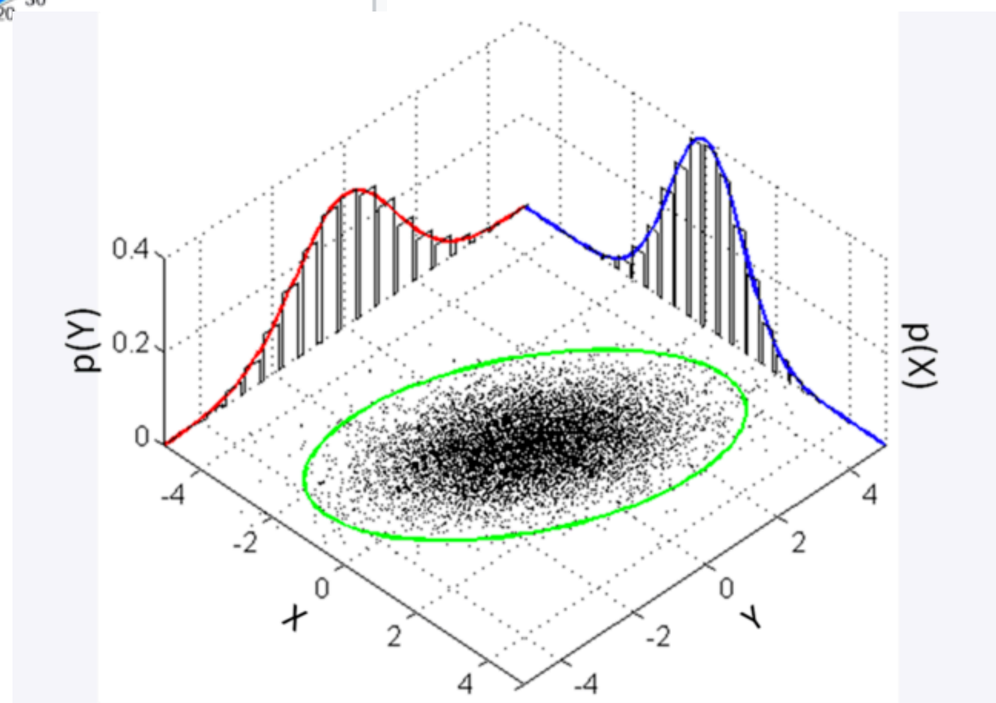
$$\boxed{f_{X,Y}(x,y) = \frac{\partial^2 F_{X,Y}(x,y)}{\partial x \partial y} \quad \textbf{(Eq.5)}}$$

Joint PDF

$$f(x,y) = \frac{\exp(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu))}{2\pi\sqrt{|\Sigma|}}$$

Bivariate normal





Bivariate normal

If X and Y have mean μ_X and μ_Y, general case is:

$f_{X,Y}(x,y) =$

$$= \frac{1}{2\pi\sigma_X\sigma_Y(1-\rho^2)^{1/2}} \exp\left[\frac{-1}{2(1-\rho^2)}\left(\frac{(x-\mu_X)^2}{\sigma_X^2} + \frac{(y-\mu_Y)^2}{\sigma_Y^2} - 2\rho\frac{(x-\mu_X)}{\sigma_X}\frac{(y-\mu_Y)}{\sigma_Y}\right)\right]$$

If X and Y are uncorrelated ($\rho = 0$), and centered with mean 0:

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y} e^{-\frac{x^2}{2\sigma_X^2} - \frac{y^2}{2\sigma_Y^2}},$$