Introduction to Statistical Machine Learning
CSC/DSCC 265/465

Lecture 4: Probability Review and Supervised Learning – Part I

Cantay Caliskan

Notes and updates

Notes and updates

- The deadline for the 1st Problem Set is Friday, January 28, 11:59 PM!
- Please send your results electronically!
- Use **Slack** for discussion
- Questions?

Plan for today

- ***Probability review***
- ***ML basics***
- ***Supervised Learning: Multivariate linear regression***
- **Direct Solution for Linear Regression**
- ***Gradient Descent Solution for Linear Regression***

Plan for today

- ***Probability review***
- *ML basics*
- *Supervised Learning: Multivariate linear regression*
- Direct Solution for Linear Regression
- *Gradient Descent Solution for Linear Regression*

Probability Review

Elements of probability

1) Sample space Ω : The set of all the outcomes of a random experiment.

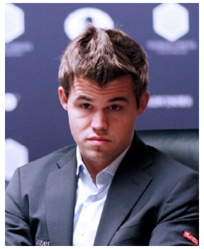
- In the sample space, each outcome $\boldsymbol{\omega} \in \boldsymbol{\Omega}$ can be thought of as a complete description of the state of the real world at the end of the experiment
- Example: We are evaluating the possible outcomes of a chess game (***win***, ***loss***, ***tie***). We should be able to associate each outcome with a probability.

2) Set of events (or **event space**) $\mathcal{F}$: A set of events in in the sample space

- Example: ***win***, ***loss***, ***tie*** (in a game of chess)

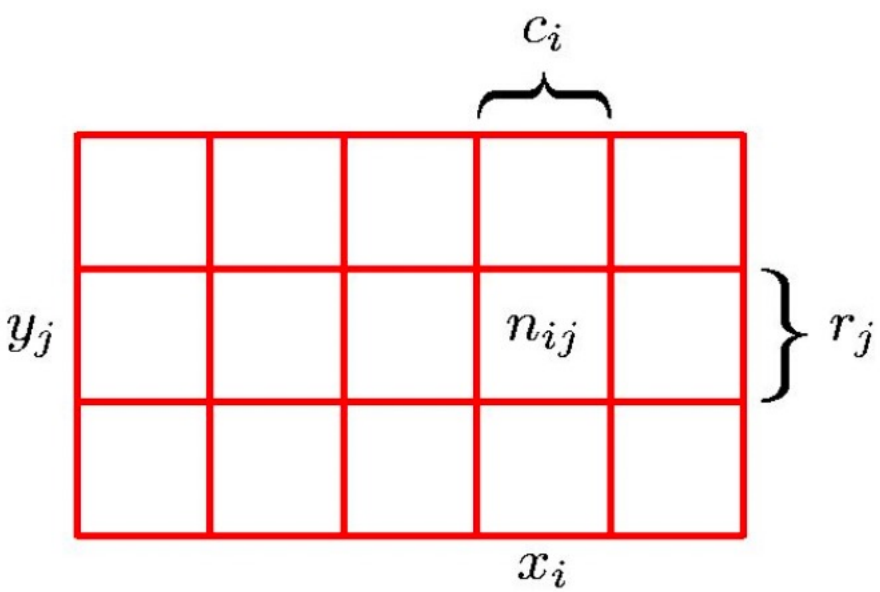
3) Probability function : Assigns each event in the event space a probability

- Example: ***function f*** calculates the winning probability of a player
  - ***f(Magnus Carlsen)*** = 0.91
  - ***f(Gary Kasparov)*** = 0.85



Question: What is min. and max. for probability function?

Definitions



Marginal Probability

$$p(X = x_i) = \frac{c_i}{N}.$$

Joint Probability

$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N}$$

Conditional Probability

$$p(Y = y_j | X = x_i) = \frac{n_{ij}}{c_i}$$

Definitions

- ***Marginal probability***: The probability of an event irrespective of the outcome of another variable
- ***Conditional probability***: The probability of one event occurring in the presence of a second event
- ***Joint probability***: The probability of two events occurring simultaneously
- Let’s take a look at an example!

UNIVERSITY of ROCHESTER

Example: Drake Equation

SOMEWHAT CERTAIN ⟷ EXTREMELY UNCERTAIN

$$N = R^{*} \times f_p \times n_e \times f_l \times f_i \times f_c \times L$$

- N: Number of **technologically advanced civilizations** in the Milky Way galaxy
- R^{*}: Rate of formation of **stars** in the galaxy
- f_p: Fraction of those stars with **planetary systems**
- n_e: Number of planets, per solar system, with an **environment suitable for life**
- f_l: Fraction of suitable planets on which **life actually appears**
- f_i: Fraction of life-bearing planets on which **intelligent life** emerges
- f_c: Fraction of civilizations that develop a **technology that releases detectable signs** of their existence into space
- L: **Length of time** such civilizations release detectable signals into space

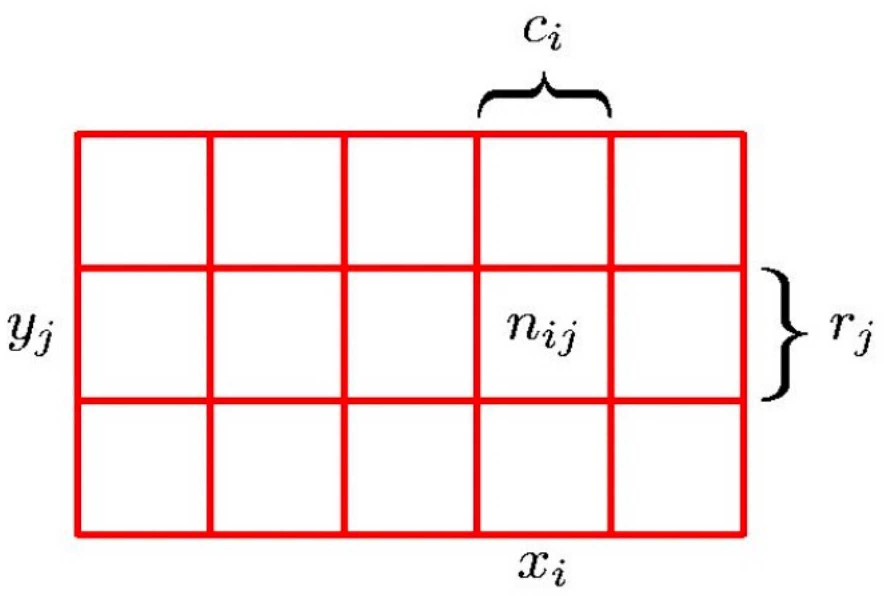
BUSINESS INSIDER

UNIVERSITY of ROCHESTER

Drake Equation

- ***Marginal probability***:
  - **P(Having a star with a planetary system)**
- ***Conditional probability:***
  - **P(Civilization with advanced technology|Life has developed)**
- ***Joint probability:***
  - **P(Environment suitable for life & Intelligent life)**

Probability Theory



Sum Rule

$$p(X = x_i) = \frac{c_i}{N} = \frac{1}{N}\sum_{j=1}^{L} n_{ij}$$

$$= \sum_{j=1}^{L} p(X = x_i, Y = y_j)$$

Product Rule

$$\begin{aligned} p(X = x_i, Y = y_j) &= \frac{n_{ij}}{N} = \frac{n_{ij}}{c_i} \cdot \frac{c_i}{N} \\ &= p(Y = y_j | X = x_i) p(X = x_i) \end{aligned}$$

Drake Equation

- ***Sum rule***
  - **P(Life in universe) = P(Life on Earth) + P(Life on Titan) + P(Life elsewhere) + …**

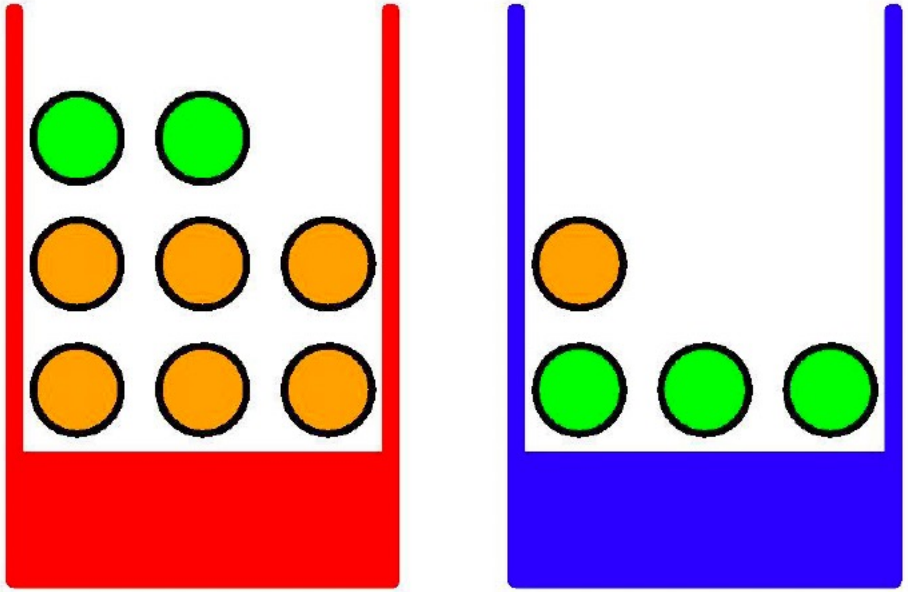
$$p(X) = \sum_Y p(X, Y)$$

- ***Product rule***
  - **P(Life on Earth, life on Titan)**

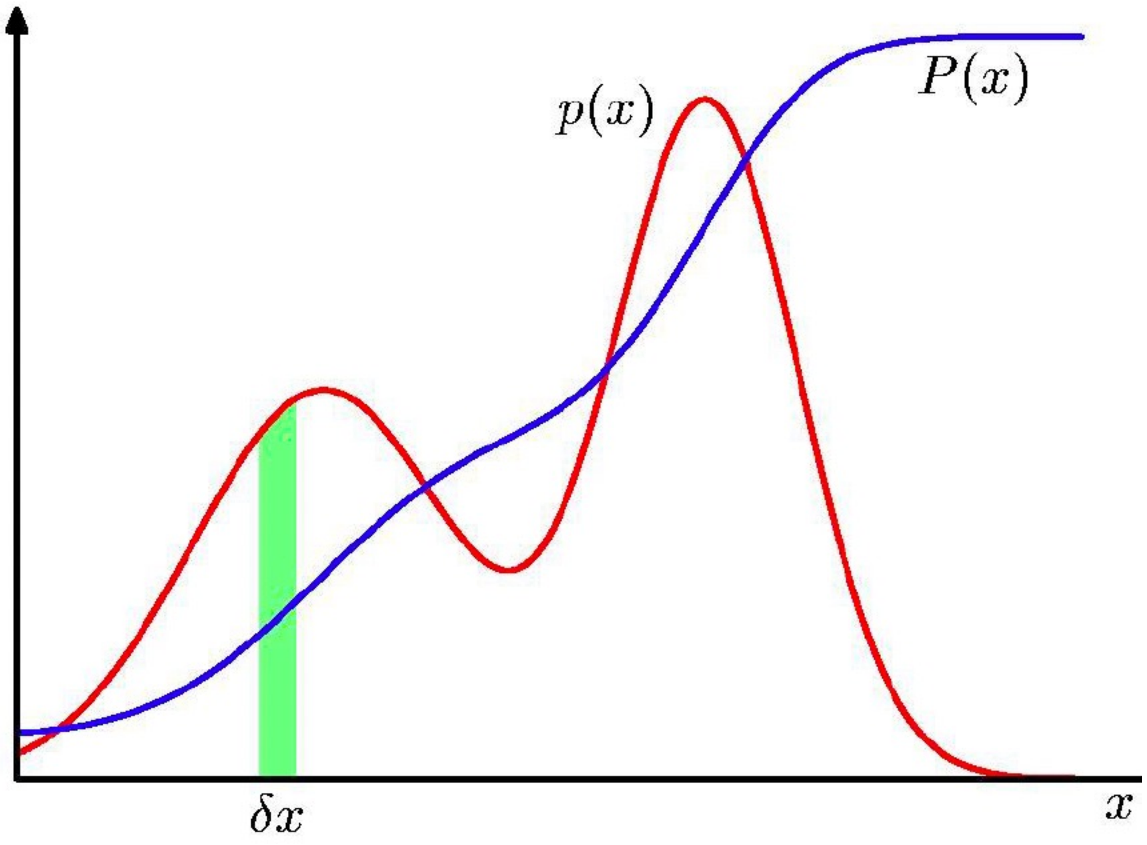
$$p(X, Y) = p(Y|X)p(X)$$

A simpler example

- Pick a random box
- Pick a random fruit
- Observe the fruit type (orange or apple)
- Put it back in the box
- Repeat the trial many times
- What is the probability of picking an apple?



Probability density function (for continuous variables)



$$p(x \in (a, b)) = \int_a^b p(x)\,\mathrm{d}x$$

$$P(z) = \int_{-\infty}^{z} p(x)\,\mathrm{d}x$$

$$p(x) \geqslant 0 \qquad \int_{-\infty}^{\infty} p(x)\,\mathrm{d}x = 1$$

Schools of Statistics
Bayesian vs. *Frequentist*

Bayesian Statistics

- Do you remember the *Bayes' formula*?

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

- Applying this formula to our hypothesis (H) and data (D):

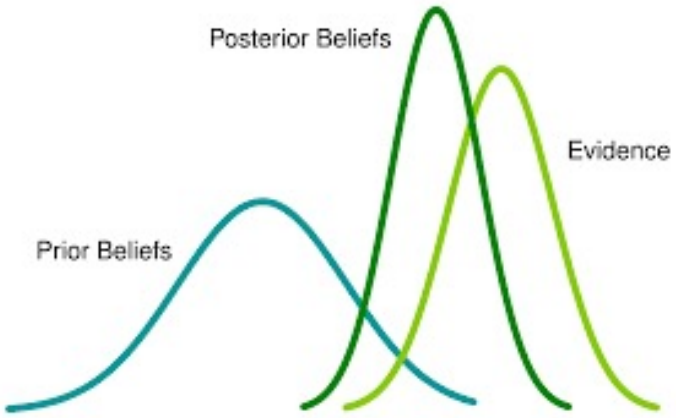
$$P(\mathcal{H} \mid \mathcal{D}) = \frac{P(\mathcal{D} \mid \mathcal{H})P(\mathcal{H})}{P(\mathcal{D})}$$

Thomas Bayes

- What do we say here?
  - We have a set of beliefs about our theory (this is our **prior**)
  - We have a dataset and we extract information from it (**likelihood**)
  - We have a new set of beliefs (this is our **posterior**)

Bayesian Statistics

$$P(\mathcal{H} \mid \mathcal{D}) = \frac{P(\mathcal{D} \mid \mathcal{H}) P(\mathcal{H})}{P(\mathcal{D})}$$



- The **prior** ***P(H)*** is the probability that ***H*** is true before the data is considered
- The **posterior** ***P(H|D)*** is the probability that ***H*** is true after the data is considered
- The **likelihood** ***P(D|H)*** is the evidence about ***H*** provided by the data ***D***
- ***P(D)*** is the **total probability** of the data taking into account all possible hypotheses

Example: Cleveland Cavaliers

- In 2020-2021 season, Cleveland Cavaliers have won **22 games** out of **72 games**
  - This data allows us to make guesses about the probability that Cleveland Cavaliers wins a game in the next season
  - The simplest guess: the winning percentage is **0.306**
  - This is the **likelihood**
  - We actually estimated **MLE** here…
- Let's expand our knowledge:
  - The winning percentages for Cleveland Cavaliers for the past five seasons were: **[0.695, 0.622, 0.610, 0.232, 0.306]**
  - The *average* winning percentage is **0.493** ← This is the **prior**
- What would be our best guess now?
  - Possibly somewhere between **0.306** and **0.493** ← This is posterior probability! (**MAP**)

Bayesian school vs. Frequentist school

- Prior: A set of beliefs about the "distribution" of the data

- *Bayesian* school:
  - Models uncertainty by a probability distribution over hypotheses. One's ability to make inferences depends on one's confidence in the prior beliefs about theory
  - Beliefs are <u>not fixed</u> and and they are subject to **Bayesian updating**

- *Frequentist* school:
  - Some hypothesis is true and the observed data is sampled from that distribution
  - Beliefs are <u>fixed</u>

Expectations

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

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

$$\mathbb{E}_x[f|y] = \sum_x p(x|y) f(x)$$

Conditional Expectation
(discrete)

$$\mathbb{E}[f] \simeq \frac{1}{N} \sum_{n=1}^{N} f(x_n)$$

Approximate Expectation
(discrete and continuous)

UNIVERSITY of ROCHESTER

Variances and Covariances

$$\text{var}[f] = \mathbb{E}\left[(f(x) - \mathbb{E}[f(x)])^2\right] = \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2$$

$$\begin{aligned} \text{cov}[x, y] &= \mathbb{E}_{x,y}\left[\{x - \mathbb{E}[x]\}\{y - \mathbb{E}[y]\}\right] \\ &= \mathbb{E}_{x,y}[xy] - \mathbb{E}[x]\mathbb{E}[y] \end{aligned}$$

Scalar form

$$\begin{aligned} \text{cov}[\mathbf{x}, \mathbf{y}] &= \mathbb{E}_{\mathbf{x},\mathbf{y}}\left[\{\mathbf{x} - \mathbb{E}[\mathbf{x}]\}\{\mathbf{y}^{\mathrm{T}} - \mathbb{E}[\mathbf{y}^{\mathrm{T}}]\}\right] \\ &= \mathbb{E}_{\mathbf{x},\mathbf{y}}[\mathbf{x}\mathbf{y}^{\mathrm{T}}] - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{y}^{\mathrm{T}}] \end{aligned}$$

Matrix form

Question: What is Cov(X,X)?

Variance-Covariance (Var-Cov) Matrix

- Usually called ***variance-covariance*** matrix
  - But, sometimes also called covariance matrix
  - **Square, symmetric** and **positive semi-definite**

- <u>Definition</u>: Let's say we have a (column) vector $\boldsymbol{x}^{\boldsymbol{T}} = [x_1\ x_2\ \ \dots\ \ x_n]$
- ***Covariance***: $cov(X) = E[(\mathrm{X} - E[\mathrm{X}])(X - E[X])^T]$ (vector space)
- ***Variance-covariance matrix***: $cov(\boldsymbol{X}) = \mathrm{E}[(\mathbf{X} - \mathrm{E}[\mathbf{X}])(\mathbf{X} - \mathrm{E}[\mathbf{X}])^T]$ (matrix space)
  - <u>Note</u>: *X* stands for vector space, and ***X*** stands for matrix space

$$\begin{bmatrix} Var(x_1) & \dots & Cov(x_1, x_n) \\ \vdots & \ddots & \vdots \\ \vdots & & \vdots \\ Cov(x_n, x_1) & \dots & Var(x_n) \end{bmatrix}$$

Sparse format

$$\begin{bmatrix} var(x) & cov(x, y) \\ cov(x, y) & var(y) \end{bmatrix}$$

2 x 2

$$\begin{bmatrix} var(x) & cov(x, y) & cov(x, z) \\ cov(x, y) & var(y) & cov(y, z) \\ cov(x, z) & cov(y, z) & var(z) \end{bmatrix}$$

3 x 3

UNIVERSITY of ROCHESTER

Definiteness

- When is a matrix definite?
  - Let's say we have a nonzero real-valued column vector $\boldsymbol{x}$
  - ***Positive-definite***

    $$M \text{ positive-definite} \iff \mathbf{x}^\mathsf{T} M \mathbf{x} > 0 \text{ for all } \mathbf{x} \in \mathbb{R}^n \setminus \{\mathbf{0}\}$$

  - ***Positive semi-definite***

    $$M \text{ positive semi-definite} \iff \mathbf{x}^\mathsf{T} M \mathbf{x} \geq 0 \text{ for all } \mathbf{x} \in \mathbb{R}^n$$

  - ***Negative definite***

    $$M \text{ negative-definite} \iff \mathbf{x}^\mathsf{T} M \mathbf{x} < 0 \text{ for all } \mathbf{x} \in \mathbb{R}^n \setminus \{\mathbf{0}\}$$

  - ***Negative semi-definite***

    $$M \text{ negative semi-definite} \iff \mathbf{x}^\mathsf{T} M \mathbf{x} \leq 0 \text{ for all } \mathbf{x} \in \mathbb{R}^n$$

- Question: How is this helpful?

Definiteness

- How is matrix definiteness helpful?

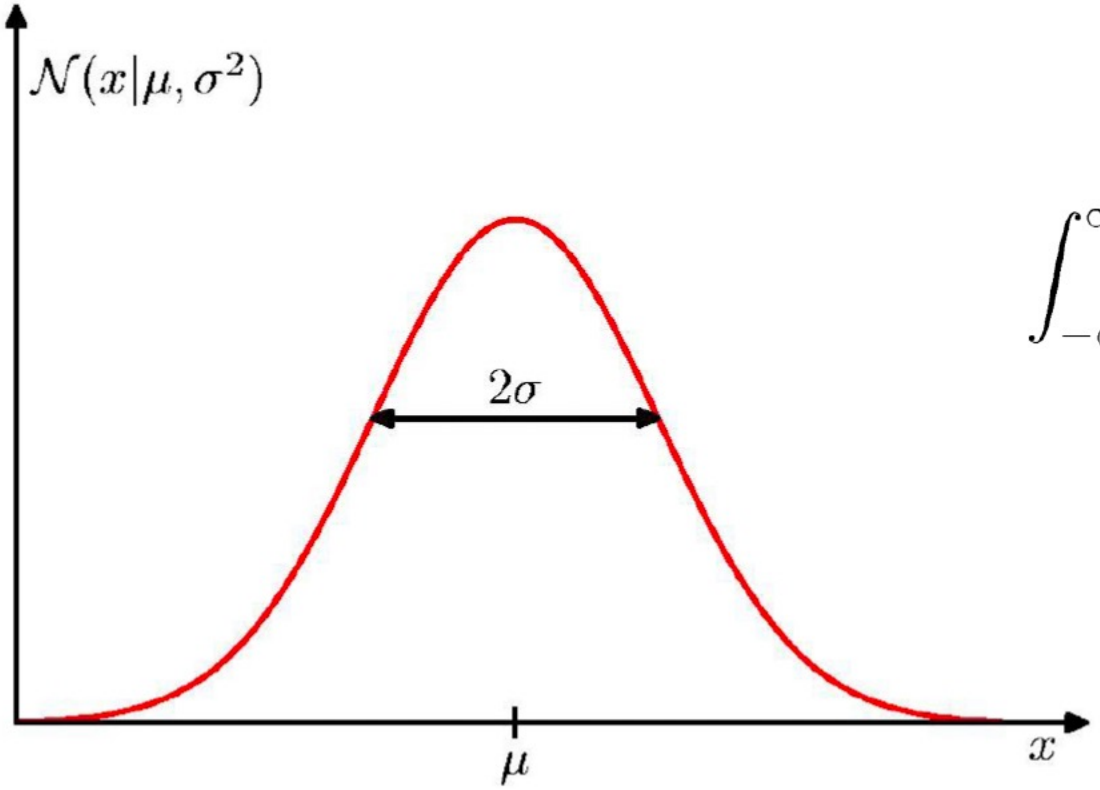
  - ***Decomposition*** or ***Factorization***:
    - **Cholesky** decomposition
    - **LDL** decomposition
      - Idea: Transform a 'not-very-manageable' matrix into manageable (smaller or more sparse matrices)
      - Quite important in graph theory!
        - Network distances, network clustering, network kernels

  - Check Jure Leskovec's class on ***Machine Learning with Graphs***: https://www.youtube.com/watch?v=JAB_plj2rbA&ab_channel=stanfordonline

The Gaussian Distribution

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

Also called normal distribution or Bell-Curve



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

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

The Gaussian Distribution

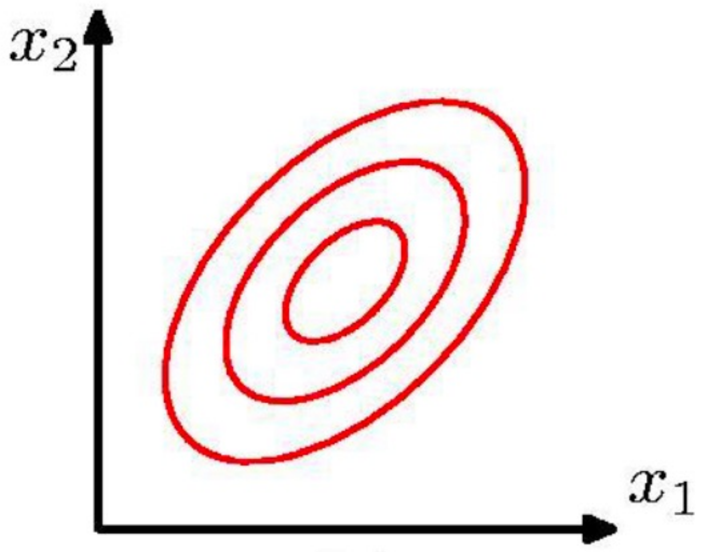
$$\mathbb{E}[x] = \int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu, \sigma^2\right) x \,\mathrm{d}x = \mu$$

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

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

The Multivariate Gaussian

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



Plan for today

- *Probability review*
- ***ML basics***
- *Supervised Learning: Multivariate linear regression*
- Direct Solution for Linear Regression
- *Gradient Descent Solution for Linear Regression*

ML Basics

Cross-Validation

- Definition: Different model validation techniques for assessing how the results of a statistical analysis (model) will generalize to an independent data set

- Usually used in the context of ***prediction***

- Why is it helpful? What is the goal?
  - Helps us to evaluate the quality of the model
  - Helps us to select the model which will perform best on unseen data
  - Helps us to avoid ***overfitting*** and ***underfitting***
  - Helps us to have a model that is low on ***bias*** and ***variance***

Bias and Variance

- Goal: Hitting the target!

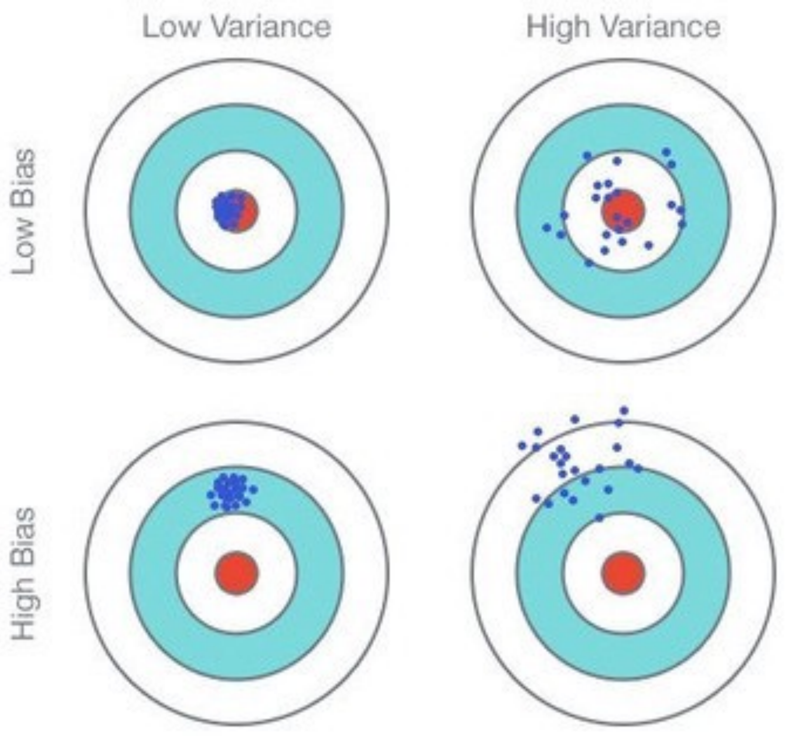


Fig. 1: Graphical Illustration of bias-variance trade-off , Source: Scott Fortmann-Roe., Understanding Bias-Variance Trade-off

Cross-Validation

- Idea:
  - You have a ***sample dataset***
    - This dataset represents the characteristics in the ***whole population***
  - This sample dataset may have some "random" differences from the population data
  - Goal: To create an explanatory / predictive algorithm from the sample dataset
    - The algorithm will almost always perform less well when applied on the whole population
    - Why? Because of random error …
  - You do cross-validation -> To reduce the ***bias*** resulting from the random errors as much as possible

Training, Validation and Testing

- **<u>Training Dataset</u>**: The sample of data used to fit the model (”train” the model)
    - During the ***training phase***, model parameters are estimated according to an optimization mechanism (according to some ***cost function***)

- **<u>Validation Dataset</u>**: The sample of data used to provide an unbiased evaluation of the model while tuning model hyperparameters

- **<u>Test Dataset</u>**: The sample of data used to provide an unbiased evaluation of a final model fit on the training dataset
    - Used to test the performance of competing models

Underfitting and Overfitting

- Underfitting: Not capturing enough patterns in the data. The model performs poorly both in the training and the test set

- Overfitting: a) Capturing noise and b) Capturing patterns which do not generalize well to unseen data. The model performs extremely well to the training set, but poorly on the test set.

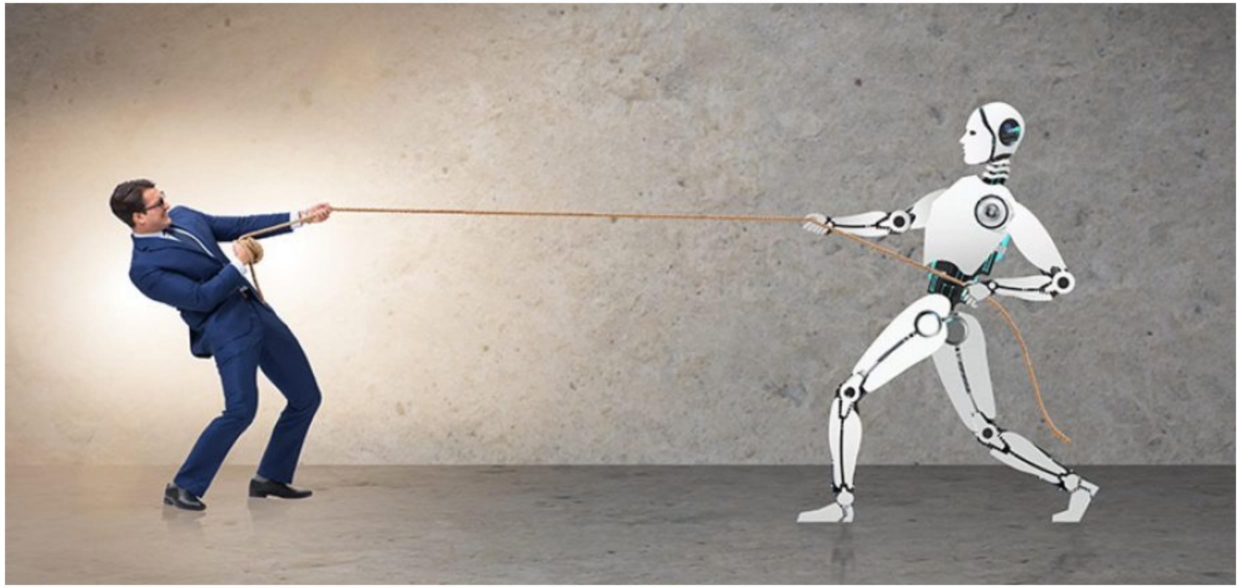
Underfitting and Overfitting

- Error on the dataset used to *fit* the model can be misleading
  - Doesn't predict future performance.
- Too much complexity can diminish model's accuracy on future data
  - Sometimes called the ***Bias-Variance*** tradeoff

Plan for today

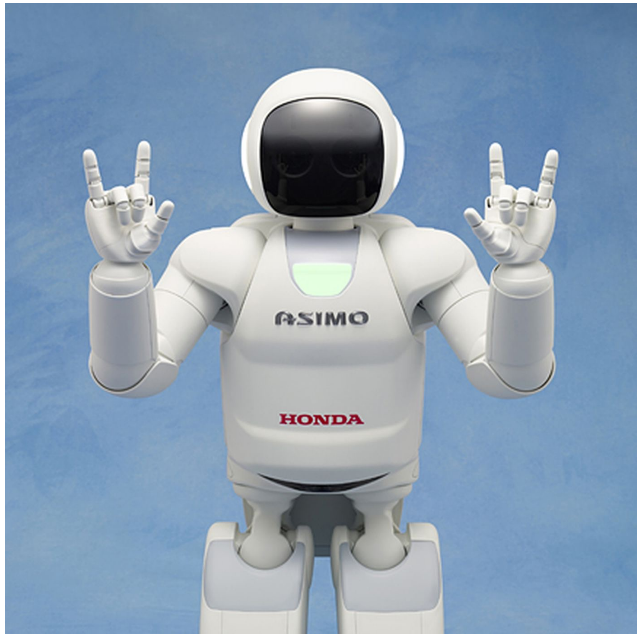
- *Probability review*
- *ML basics*
- ***Supervised Learning: Multivariate linear regression***
- Direct Solution for Linear Regression
- *Gradient Descent Solution for Linear Regression*

Supervised Learning



Supervised Learning

- Provide a ***training set*** to your algorithm
  - Your set needs to have ***features*** and ***labels***
- Your training performance is tested with a ***test set***
- Outputs can be:
  - Categorical (***classification***)
  - Continuous (***regression***)



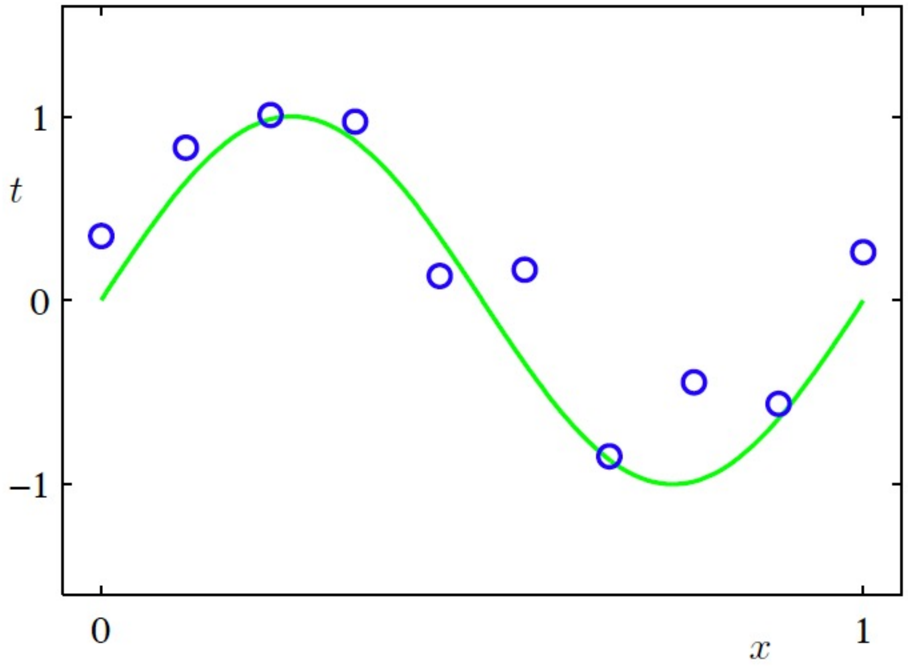
Examples: Supervised Learning

- **Recognize digits**
  - MNIST dataset!
  - What is input, what is output?

- **Predict the future prices of Tesla stock**
  - Yahoo Finance!
  - What is input, what is output?

Curve Fitting

- A training dataset with **N = 10** (samples) shown as **blue circles**
  - Each have an input (or series of input) ***x***
  - And a target variable ***t***



<u>Goal</u>: Predict the value ***t*** for some new ***x***

<u>Basic form:</u>

- $y(\boldsymbol{x}, \boldsymbol{w}) = w_0 + w_1 x_1 + \ldots + w_d x_d$

d : Dimension

- $x = (x_1, \ldots, x_d)^T$ is the ***input vector***
- $w = (w_1, \ldots, w_d)^T$ is the ***weight vector*** (parameters to be estimated)

Linear Regression

- More general form of linear regression:

$$y(\boldsymbol{x}, \boldsymbol{w}) = w_0 + \sum_{j=1}^{M-1} w_j \, \phi_j(\boldsymbol{x})$$

- where $\phi_j(\boldsymbol{x})$'s are called basis functions

- Parameter w_0 is called a ***basis parameter*** or a ***constant term***.
- If we add $\phi_0(\boldsymbol{x}) = 1$, we get:

$$y(\boldsymbol{x}, \boldsymbol{w}) = \sum_{j=1}^{M-1} \boldsymbol{w}^T \, \phi_j(\boldsymbol{x})$$

Quick Note: Basis Functions

- **Polynomial functions:**

$$\phi_j(x) = x^j$$

- **Gaussian functions:**

$$\phi_j(\mathbf{x}) = \exp\left\{-\frac{(\mathbf{x}-\boldsymbol{\mu}_j)^T(\mathbf{x}-\boldsymbol{\mu}_j)}{2s^2}\right\}$$

- **Sigmoid basis functions:**

$$\phi_j(\mathbf{x}) = \sigma\left(\frac{\mathbf{b}^T\mathbf{x}-\mu_j}{s}\right)$$

- **Wavelets:**
  - Continuous wavelet transform, fractional Fourier transform etc.

<u>Question</u>: Why do we need different basis functions?

Please do the following until next lecture!

- Review what we have just gone through (supervised learning: linaer regression)
- Continue with your problem set (due date ***Friday, January 28, 11:59 PM***)
- Read ***Chapter 4*** from our book