香港中文大學（深圳）
The Chinese University of Hong Kong, Shenzhen

Introduction to Data Science

# Lecture 24 Review for Final Exam

Zicheng Wang

- Final Exam: May 15th, 8:30-10:30am, Location TBA
- Zero tolerance policy when it comes to cheating so, DON’T CHEAT
- Missing the final exam without prior notification to and approval of the instructors will automatically result in the "0" grade for the exam.
- You can bring one calculator, one non-electronic dictionary, one sheet of A4 paper with notes. Please note that any content on this sheet must be hand-written by yourself.

# Probability

# Probability Terminologies

- **Random Experiment**: a repeatable procedure
- **Sample space**: set of all possible outcomes Ω.
- **Event**: a subset of the sample space.
- **Probability mass function (discrete)**, $P(\omega)$: gives the probability for each outcome $\omega \in \Omega$
- **Probability density function (continuous)**, $f(x)$: $P(a \leq X \leq b) = \int_a^b f(x)dx$

# Probability Distributions

- Bernoulli Distribution: take value 1 with probability $p$ and value 0 with probability $1 - p$. Mean$= p$. Variance$= p(1 - p)$.

- Binomial Distribution: the number of success in $N$ independent experiments (for each experiment, the probability of success is $p$). Mean$= Np$. Variance$=Np(1 - p)$.

- Geometric Distribution: the number of experiments needed to get the first success (for each experiment, the probability of success is $p$). Mean$= 1/p$. Variance$= (1 - p)/p^2$.

# Probability Distributions

- Uniform Distribution: with the same likelihood, $X$ takes a value within $[a, b]$ where $b > a$. $f(x) = 1/(b-a)$ for $x \in [a, b]$, and $f(x) = 0$ otherwise. Mean$= (a+b)/2$. Variance$= (b-a)^2/12$.

- Normal Distribution: Mean$= \mu$. Variance$= \sigma^2$.

$$f(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)。$$

# Cumulative Distribution Function

- The CDF of a discrete random variable $X$ is

$$F(x) = P(X \leq x) = \sum_{\tilde{x} \leq x} f(\tilde{x})$$

- CDF for continuous random variable is defined as:

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(u)du$$

$$\left.\begin{array}{l} \checkmark\ 0 \leq F(x) \leq 1 \\ \checkmark\ \text{If } x \leq y, \text{ then } F(x) \leq F(y) \end{array}\right\} \text{For both discrete and continuous RVs}$$

# Mean and Variance

- Discrete:
  - ✓ Probability mass function.
- Continuous
  - ✓ Probability density function.

**Summation ↔ Integration**

- Mean

  E$[X]= \sum x\, f(x)$

- Variance

  Var$[X]= \sum (x - \mathrm{E}[X])^2\, f(x)$

- Mean

  E$[X]= \int_{-\infty}^{\infty} x f(x) dx$

- Variance

  Var$[X]= \int_{-\infty}^{\infty} \left(x - E(X)\right)^2 f(x) dx$

# Expectation

- $E\left[\sum_i C_i X_i\right] = \sum_i C_i E[X_i]$

- $E[C] = C$

- $E[g(X)] = \int_{-\infty}^{\infty} g(x) f(x) dx$

# Conditional Probability

- Given the realization of event A, the probability of event B may change

- (Conditional probability) If A and B are events with P(B)>0, then the conditional probability of A given B, denoted by P(A|B), is defined as $P(A|B) = \frac{P(A \cap B)}{P(B)}$

- (Independence) Two events A and B are called independent if and only if $P(A \cap B) = \mathrm{P}(A)P(B)$ $\quad P(A|B) = \frac{P(A \cap B)}{P(B)} = \mathrm{P(A)}$

# Statistics

# Maximum Likelihood Estimate

- Given a model with an unknown parameter $\theta$
- Given samples: $X_1,\ X_2 \ldots,\ X_n$
- The probability that the models generates the samples is called **likelihood**. L($\theta$)= P ($X_1,\ X_2 ,\ldots,\ X_n | \theta$)
- To determine the best $\theta \in \Theta$, we choose $\hat{\theta}$ such that L($\theta$) is maximized at $\theta = \hat{\theta}$. **Maximum likelihood estimate (MLE)**

# Likelihood Function

- Given a model with an unknown parameter $\theta$
- Given samples: $X_1,\ X_2 ,\ldots,\ X_n$

Continuous RV model:

- Likelihood: L($\theta$)= $\Pi_i$ f($X_i\,|\,\theta$)
- Log-Likelihood: $l(\theta)$= $\sum_i \log(\mathrm{f}(X_i\,|\,\theta))$

f: the model's probability density function (PDF)

Discrete RV model:

- Likelihood: L($\theta$)= $\Pi_i$P ($X_i\,|\,\theta$)
- Log-Likelihood: $l(\theta)$= $\sum_i \log(\mathrm{P}(X_i\,|\,\theta))$

P: the model's probability mass function (PMF)

# Example

- Suppose heights of students follow a normal distribution.
  - Given parameter $\mu$, the model is $N(\mu, 1)$

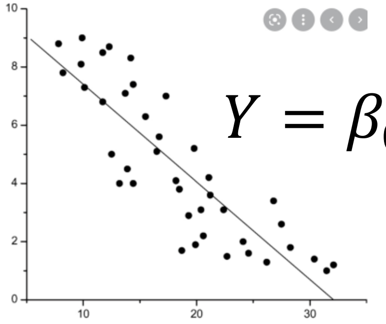
- Samples: $X_1, X_2, \ldots, X_n$

- L($\mu$) = $\left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\sum_i (X_i - \mu)^2}$
- $l(\mu)$ = Log L($\mu$) = $-\frac{1}{2}\sum_i (X_i - \mu)^2 - \frac{\mathrm{n}}{2}\log 2\pi$

- $l'(\mu)$ = $\sum_i (X_i - \mu)$= 0 $\Longrightarrow$ $\hat{\mu} = \bar{X}$.

Hint: Go through all homework problems in Hmw 2, Lecture 11

# Linear Regression




$$Y = \beta_0 + \beta_1 X$$

- $Y \sim N(\beta_0 + \beta_1 X, \sigma^2)$
  - Given the observation of X
  - Y follows a normal distribution with mean $\beta_0 + \beta_1 X$

- Regression analysis: knowing $\underbrace{\beta_0, \beta_1}$, $\sigma^2$, you can predict $X$ given $Y$

MLE: choose the best $\beta_0$, $\beta_1$

$$\text{MLE:} \begin{cases} \widehat{\beta_1} = \dfrac{\Sigma_i (X_i - \bar{X})\,(Y_i - \bar{Y})}{\Sigma_i (X_i - \bar{X})^2} \\[2ex] \widehat{\beta_0} = \bar{Y} - \widehat{\beta_1}\bar{X} \end{cases}$$

# Optimization

# Optimization problem in standard form

Objective Function

$$\begin{array}{ll} \text{minimize} & \boxed{f_0(x)} \\ \text{subject to} & \boxed{\begin{array}{l} f_i(x) \le 0, \\ h_i(x) = 0, \end{array}} \quad \begin{array}{l} i = 1, \dots, m \\ i = 1, \dots, p \end{array} \end{array}$$

Decision Variables

Constraints

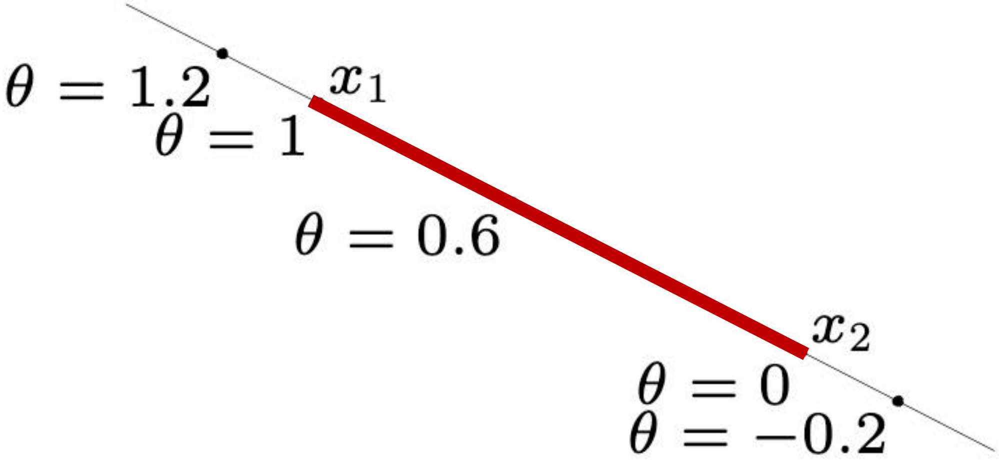
- $\boxed{x} \in \mathbf{R}^n$ is the optimization variable
- $f_0 : \mathbf{R}^n \to \mathbf{R}$ is the objective or cost function
- $f_i : \mathbf{R}^n \to \mathbf{R}$, $i = 1, \dots, m$, are the inequality constraint functions
- $h_i : \mathbf{R}^n \to \mathbf{R}$ are the equality constraint functions

# Line Segment

- Let $\boldsymbol{x_1} \neq \boldsymbol{x_2}$ be two points in $\mathbb{R}^n$. Points of the form

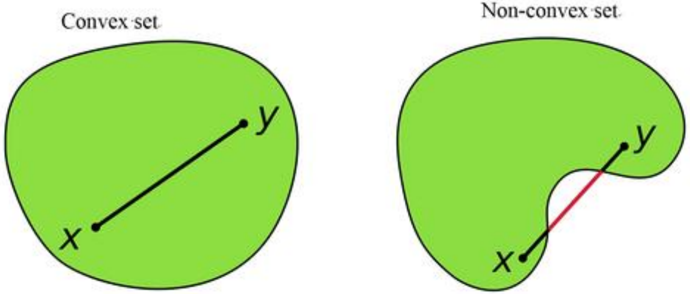
$$\boldsymbol{x} = \theta \boldsymbol{x}_1 + (1-\theta)\boldsymbol{x}_2$$

where $\boldsymbol{\theta} \in [\mathbf{0}, \mathbf{1}]$, form the **line segment** between $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$.



# Convex Set

- Set C is a **convex set** if the line segment between any two points in C lies in C.




- Formal definition: A set $C$ is convex if $\forall\ \boldsymbol{x_1}, \boldsymbol{x_2} \in C, \forall \theta \in [0,1]$
$$\theta \boldsymbol{x_1} + (1-\theta)\boldsymbol{x_2} \in C.$$

Remark: In this lecture, I will use **bold** form to represent a high dimension point. Without bold form, it represents a scalar

# Convex Set Examples

- The empty set $\emptyset$, the singleton set $\{\boldsymbol{x_0}\}$, and the complete space $R$ are convex sets.
- An interval of $[a, b] \subset R$ is a convex set
- In $R^n$ the set $\mathrm{H} \coloneqq \{\boldsymbol{x} \in R^n \colon a_1 x_1 + \cdots + a_n x_n = c\}$ is a convex set
- Half spaces, e.g., $\mathrm{H} \coloneqq \{(x, y) \colon y \leq ax + b\}$ are convex sets
- A disk with center $(0,0)$ and radius $c$ is a convex subset of $R^2$

Remark: In this lecture, I will use **bold** form to represent a high dimension point. Without bold form, it represents a scalar

# Steps for Showing the Convexity of a Set

Prove $\mathrm{H} := \{(x, y) : y = ax + b\}$ is a convex set

For any $(x_1, y_1)$ and $(x_2, y_2)$ in $H$,

- $y_1 = ax_1 + b$
- $y_2 = ax_2 + b$

1. Use the assumption that $(x_1, y_1), (x_2, \mathrm{y}_2) \in H$

- $\theta(x_1, y_1) + (1-\theta)(x_2, y_2) = \underbrace{(\theta x_1 + (1-\theta)x_2), \theta y_1 + (1-\theta)y_2)}$

- Then for any $\theta \in [0,1]$

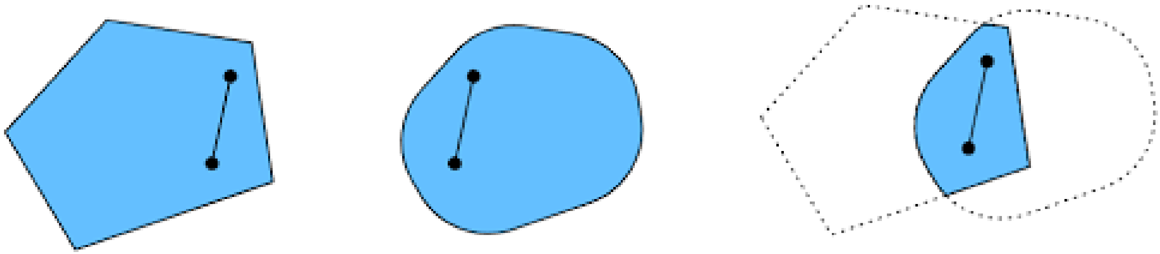
2. Characterize the new point within the line segment

  - $\underbrace{\theta y_1 + (1-\theta)y_2 = a(\theta x_1 + (1-\theta)x_2) + b}$

3. Use (1) and (2) to show that the new point is in $H$

# Property of Convex Sets

Lemma: If both $S_1$ and $S_2$ are convex sets, then $S_1 \cap S_2$ is also a convex set.



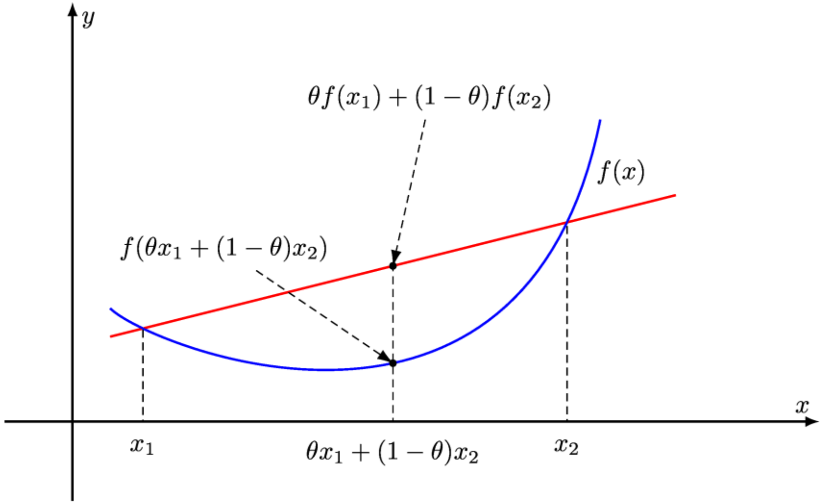
Hint: Go through homework problems 1.6, 2.2, 2.3, 2.7,.2.8 and similar problems in Hmw 3

# Convex Function

Definition: A function $f(\boldsymbol{x}): R^n \rightarrow R$ is **convex** if (1) its domain is a convex set, and (2) for any $\boldsymbol{x_1}, \boldsymbol{x_2} \in dom(f)$ and any $0 \leq \lambda \leq 1$, we have

$$f(\boldsymbol{z}) \leq \lambda f(\boldsymbol{x_1}) + (1-\lambda) f(\boldsymbol{x_2})$$

where $\boldsymbol{z} = \lambda \boldsymbol{x_1} + (1-\lambda)\boldsymbol{x_2}$.



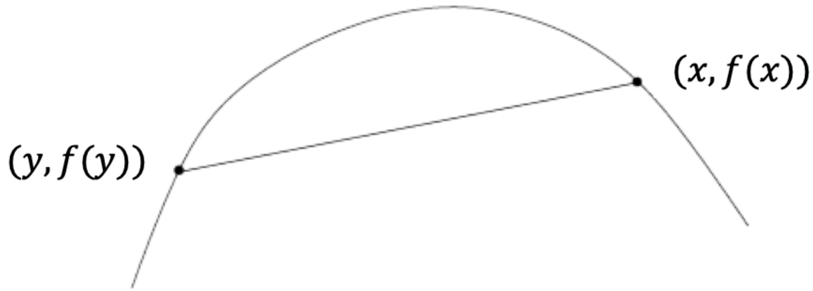

**Function $f$ evaluated at the combination** of two points $\boldsymbol{x_1}, \boldsymbol{x_2}$ is no larger than the **same combination of $f(\boldsymbol{x_1})$** and $\boldsymbol{f(x_2)}$

# Concave Function

Definition: A function $f(\boldsymbol{x})\colon R^n \to R$ is **concave** if (1) the domain of $f$ is a convex set, and (2) for any $\boldsymbol{x}, \boldsymbol{y} \in dom(f)$ and any $0 \leq \lambda \leq 1$, we have

$$f(\boldsymbol{z}) \geq \lambda f(\boldsymbol{x}) + (1-\lambda) f(\boldsymbol{y})$$

where $\boldsymbol{z} = \lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y}$.




**If f is concave, then -f is convex!**

**If f is convex, then -f is concave!**

# Second Order Condition

Suppose $f$ is a twice continuously differentiable function. Then $f$ is convex **if and only if**

(1) dom(f) is a convex set

(2) for any $\boldsymbol{x} \in$ dom(f), any unit vector $\boldsymbol{e}$ satisfying that there exists $\epsilon > 0$ such that $\boldsymbol{x} + \epsilon\boldsymbol{e} \in$ dom(f),

$$\frac{d^2 f(\boldsymbol{x} + \theta\boldsymbol{e})}{d\theta^2}(0) \geq 0$$

One dimension: $\boldsymbol{f''(x)} \geq \mathbf{0}$

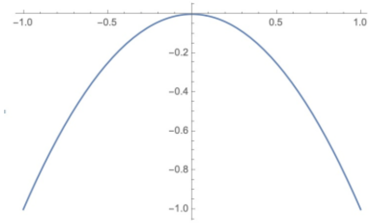
# Examples

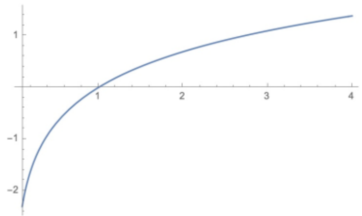
## Convex

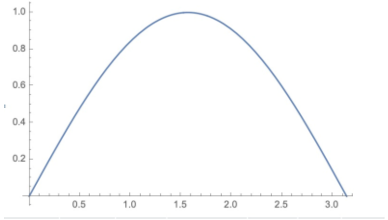
- $f(x) = ax + b\ (also\ concave)$
- $f(x) = x^2$
- $f(x) = e^x$

## Concave

- $f(x) = -x^2$
- $f(x) = \log(x)$ on $(0, +\infty)$
- $f(x) = \sin(x)$ on $[0, \pi]$



# Convex Function VS. Convex Set

- $C = \{\boldsymbol{x}: f(\boldsymbol{x}) \leq r\}$ is a convex set if $f(\boldsymbol{x})$ is a convex function
  - $C$ is also called a sublevel set of $f(\boldsymbol{x})$

- $C = \{(\boldsymbol{x}, y): y \geq f(\boldsymbol{x})\}$ is a convex set if and only if $f(\boldsymbol{x})$ is a convex function.
  - $C$ is also called the epigraph of $f(\boldsymbol{x})$

# Application

Prove a unit disk, e.g., $\mathrm{H} \coloneqq \{(x, y): x^2 + y^2 \leq 1\}$ is a convex set.

We consider $f(\boldsymbol{x}) = \sum_i a_i x_i^2$ with $a_i$>0. Given any point $\mathbf{y}$, any unit vector $\boldsymbol{e}$ and any $\theta$

$$g(\theta) = f(\boldsymbol{y} + \theta \boldsymbol{e}) = \sum_i a_i (y_i + \theta e_i)^2$$

Then g''(0) = $\sum_i 2a_i e_i^2 \geq 0$. So $f(\boldsymbol{x})$ is a convex function.

$\{\boldsymbol{x}: f(\boldsymbol{x}) \leq r\}$ forms a ball/disk or an ellipsoid, so it is a convex set.

# Operations Preserving Convexity

If $f_1, \ldots, f_m$ are convex functions, then $f(\boldsymbol{x}) = \max\{f_1(\boldsymbol{x}), \ldots, f_m(\boldsymbol{x})\}$ is also convex.

- **Maximum of a set of convex functions**

If $f_1, f_2, \ldots, f_n$ are convex functions, and $w_1, w_2, \ldots, w_n \geq 0$, then $f = w_1 f_1 + w_2 f_2 + \cdots + w_n f_n$ is also a convex function.

- **Nonnegative weighted sums of convex functions**

Hint: Go through homework problems 3.1, 3.2, 3.3, 3.6, 3.8, 5.1 and similar problems in Hmw 3

# Convex Optimization Problem

$$
\begin{aligned}
&\text{minimize} && f(x) \\
&\text{subject to} && g_i(x) \leq 0, i = 1, \dots, m \\
& && h_i(x) = 0, i = 1, \dots, n
\end{aligned}
$$

A convex optimization problem needs to satisfy the following two conditions:

- Its feasible set is a **convex set**.
- Its objective function is a **convex function**.

# Why Convex Optimization Problem?

- Any local minimum is also a global minimum.
- Any interior local minimum satisfies the first order condition.

$$\nabla f(p) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(p) \\ \vdots \\ \frac{\partial f}{\partial x_n}(p) \end{bmatrix} \longrightarrow \nabla f(\boldsymbol{x}^*) = \mathbf{0}$$

# Example: Newsvendor Problem

- Suppose you want to start your own blind box business.
- Let $D$ denote the one season (three months) random demand, which follows a uniform distribution in $[10,100]$.
- At the beginning of each season, you place an order $Q$ to Pop Mart, with a cost $10$ Yuan for each blind box.
- Each blind box can be sold at a price of $20$ Yuan.
- At the end of each season, unsold blind boxes are salvaged, and you get $3$ Yuan for each salvaged box.

- How many blind boxes should you order to maximize your expected profit?

# Machine Learning

# Supervised Learning

- Supervised machine learning algorithms utilize labeled data for training, where the correct outputs corresponding to input data are already known.

- For all samples, $(x^i, y^i), i = 1, \ldots N$, you can observe both the input data $x^i$ and the label $y^i$

Training data

y=1 (cat)

y=0 (dog)

y=1 (cat)

…… ……

y=0 (dog)

# Supervised Learning

Training data

y=1 (cat)

y=0 (dog)

y=1 (cat)

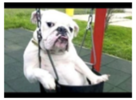
…… ……

y=0 (dog)

Learning algorithm (optimization involved)

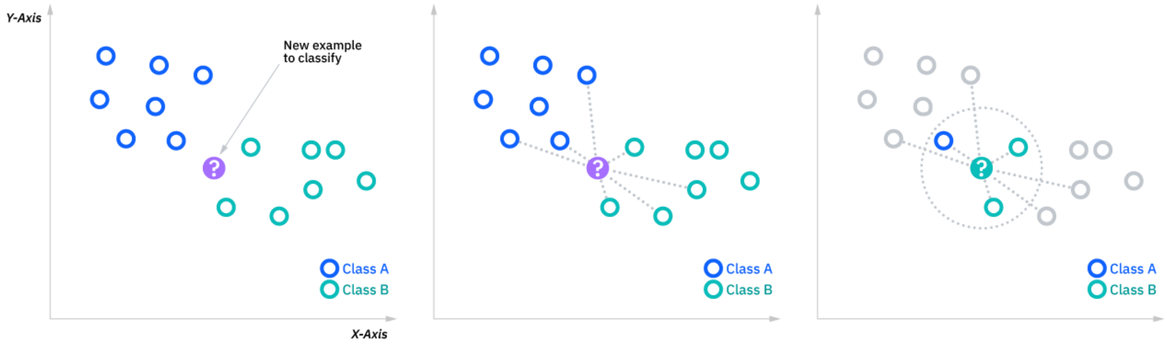
Classifier $h: X \rightarrow \{0,1\}$

For example:

$h\Big($  $\Big) \longrightarrow$ 0

# K-Nearest Neighbor Classifier

- Find $K$ training points $x_i$ closest to $x$.
- If the majority of K-nearest neighbors of $x$ belong to classifier c, label x as c.



# The KNN Algorithm

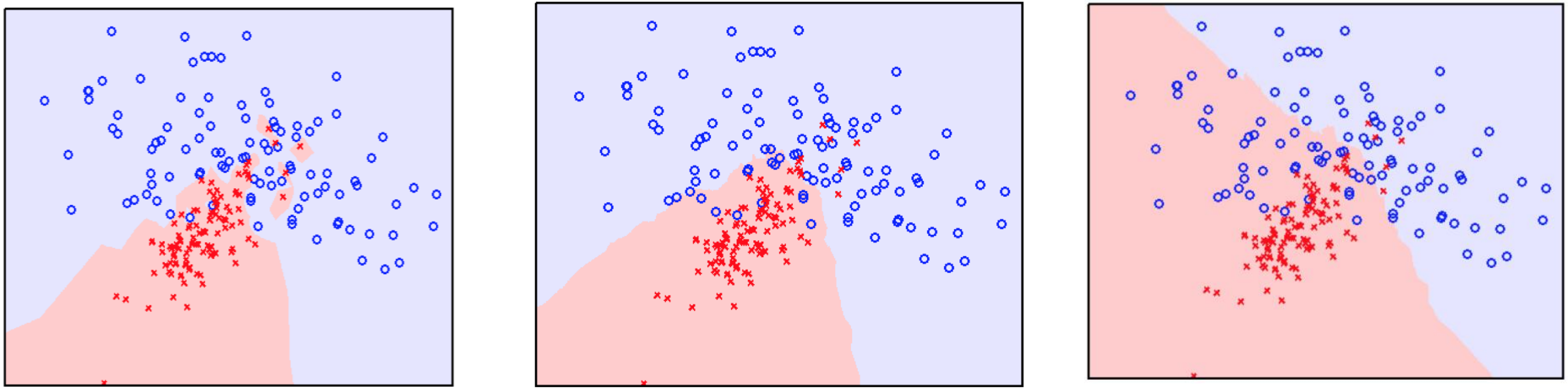
1. Load the data
2. Set $K$ of your choice to be the number of neighbors
3. For each new data to be classified
   - Calculate the distances between the new data and all the labeled data.
   - Record the entry $(d_i, y_i)$, where $d_i$ is the distance between the new data and the ith labeled data, and $y_i$ is the label of the ith data.
   - Sort the these entries with respect to distance (from smallest to largest).

5. Pick the first K entries from the sorted collection
6. Get the labels of the selected K entries
7. Choose the label with the largest frequency

# The Choice of K

Overfitting

Underfitting



$K = 1$

$K = 25$

$K = 101$

# Overfitting and Underfitting

- A machine learning model is said to **overfit** the data when it learns patterns specific to the training data and make accurate predictions only on the training data.

- A machine learning model is said to **underfit** when it fails to capture the key patterns or relationships between variables in both the training and test data.

# Logistic Regression

- Model the conditional probability of the label given the data
- **Simplest** case (two classes): y ∈ $\{0 , 1\}$
- **Logistic regression model:**

Logistic function

$$p(y = 1|\boldsymbol{x}, \boldsymbol{\theta}, b) = \boxed{\frac{1}{1 + \exp(-(\boldsymbol{\theta}^\top \boldsymbol{x} + b))}}$$

$$p(y = 0|\boldsymbol{x}, \boldsymbol{\theta}, b) = \frac{\exp(-(\boldsymbol{\theta}^\top \boldsymbol{x} + b))}{1 + \exp(-(\boldsymbol{\theta}^\top \boldsymbol{x} + b))}$$

# Train the Logistic Regression Model

- **How to find $\theta$ *and* $b$? MLE**
- Given m labeled samples $(\boldsymbol{x}^i, y^i)$, i = 1, ...m
- Find $\boldsymbol{\theta}$ and $b$ such that the likelihood of observing the labeled samples is maximized

$$\max_{\boldsymbol{\theta},b} l(\boldsymbol{\theta}, b) := \log \prod_{i=1}^{m} P\left(y^i \middle| \boldsymbol{x}^i, \boldsymbol{\theta}, b\right) = \sum_{i=1}^{m} \log P\left(y^i \middle| \boldsymbol{x}^i, \boldsymbol{\theta}, b\right)$$

- Usually, we equivalently maximize the averaged likelihood

$$\max_{\boldsymbol{\theta},b} \frac{1}{m} l(\boldsymbol{\theta}, b) := \frac{1}{m} \sum_{i=1}^{m} \log P\left(y^i \middle| \boldsymbol{x}^i, \boldsymbol{\theta}, b\right)$$

**Bad news**: no closed form solution to the problem

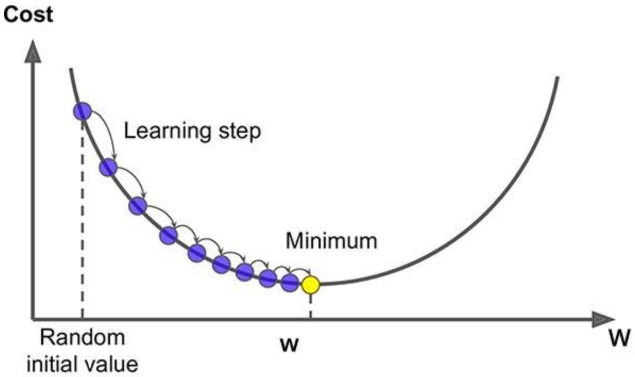
# Gradient Descent Method

- Start with an initial point $x^{(0)}$

$\alpha^{(t)}$ : the step size or learning rate

- Update our point by the following rule:

$$x^{(t+1)} = x^{(t)} - \boxed{\alpha^{(t)}} f'(x^{(t)})$$




- Stopping criteria:
  - $|x^{(t+1)} - x^{(t)}| \leq \varepsilon$
  - or $|f'(x^{(t)})| \leq \varepsilon$

How to select $\alpha^{(t)}$? The selection of $\alpha^{(t)}$ will affect the rate at which we find the local minimizer. A bad selection of $\alpha^{(t)}$ can result in the failure of the algorithm.

# If $\alpha^{(t)}$ is not well chosen, we may not meet the stop criteria.

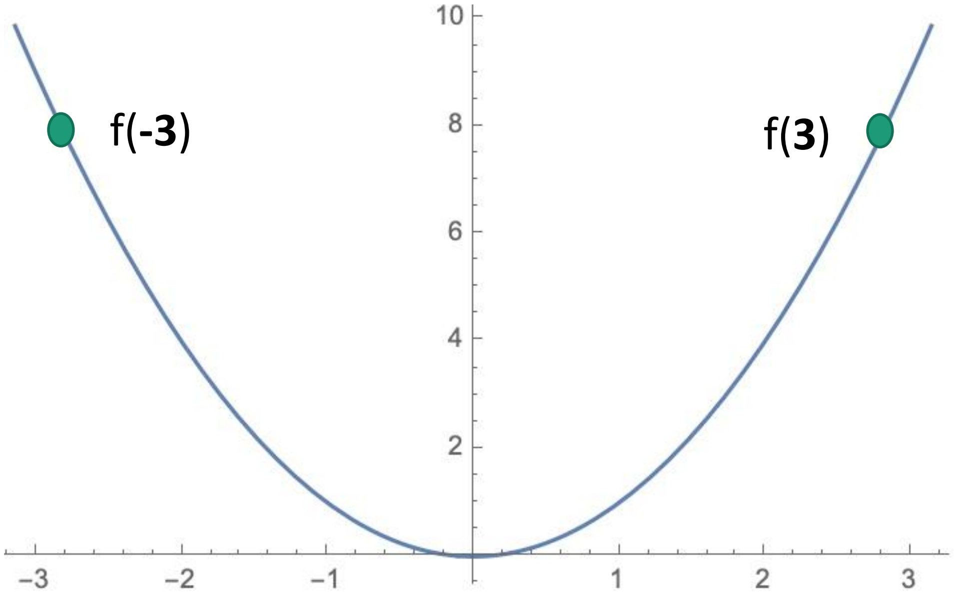
f(x) = $x^2$

Suppose $\alpha^{(t)}$ = 1.

$x^{(t+1)} = x^{(t)} - f'(x^{(t)})$

If $x^{(t)}$ = -3
- f'(-3) = -6
- $x^{(t+1)}$ = 3

If $x^{(t)}$ = 3
- f'(3) = 6
- $x^{(t+1)}$ = -3




The updated points will oscillate between 3 and -3

# Unsupervised Learning

- Data lacks structured or objective answers, such as labels.
- In other words, for all samples $(x^i, y^i)$, where $i = 1, \ldots N$, you can observe $x^i$ but $y^i$ remains unseen.

# Unsupervised Learning

- There is no predefined correct output for a given input.
- Instead, the algorithm must interpret the input and make the appropriate decision.
- The aim is to **examine the data and discern underlying patterns**.

# K-Means Clustering

- Given $m$ data points, $\{x^1, x^2, \dots x^m\}$
- Find $k$ cluster centers, $\{c^1, c^2, \dots, c^k\}$
- And assign each data point $i$ to one cluster, $\pi(i) \in \{1, \dots, k\}$
- Such that the sum of the distances from each data point to its respective cluster center is minimized

# K-Means Clustering

- Step 1: Initialize $k$ cluster centers, $\{c^1, c^2, \ldots, c^k\}$, randomly
- Step 2: Do
  - Decide the cluster memberships of each data point, $x^i$, by assigning it to the nearest cluster center (cluster assignment)

$$\pi(i) = argmin_{j=1,\ldots,k} \left\| x^i - c^j \right\|^2$$

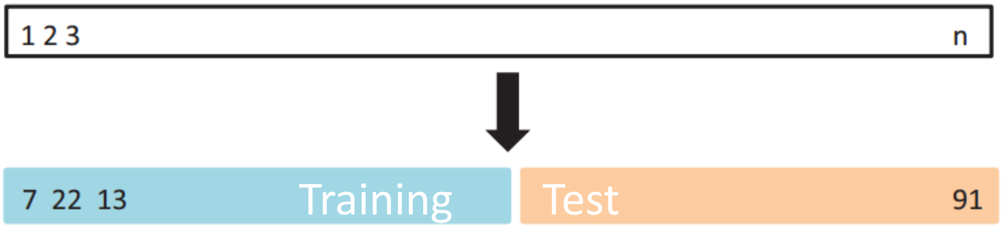
  - Adjust the cluster centers (center adjustment)

$$c^j = \frac{1}{|\{i : \pi(i) = j\}|} \sum_{i:\pi(i)=j} x^i$$

- While any cluster center undergoes changes, go to Step 2

# Model Selection: Validation Set Approach

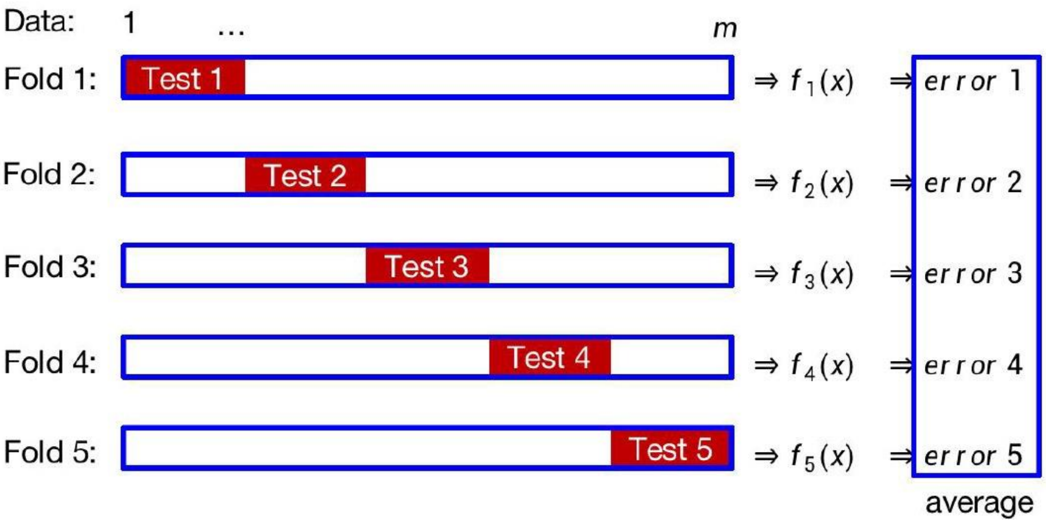
- Divide samples in to **training data** and **test data**.




- A set of model to choose {1,2,..., M}. (e.g, KNN, choose K=1,2,...M)
- For each model m,
  - use training data to train model
  - use the learned model to calculate the prediction error for test data. $Err_m$
- Choose model that has the smallest test error (min $Err_m$)

# K-Fold Validation

- 5-fold cross-validation (blank: training; red: test)




- $f_i$ is fitted by the training data in Fold i.
- For each fold, use test data to test the prediction error.
- Use the **average** error as the model's prediction error.

Error

| **Model 1** $\mathbf{h}(\boldsymbol{\theta}, \boldsymbol{X})$ | **Model 2** $\mathbf{g}(\boldsymbol{\gamma}, \boldsymbol{X})$ |
|---|---|
| $\mathrm{Err}_h(\boldsymbol{\theta}_1)$ | $\mathrm{Err}_g(\boldsymbol{\gamma}_1)$ |
| $\mathrm{Err}_h(\boldsymbol{\theta}_2)$ | $\mathrm{Err}_g(\boldsymbol{\gamma}_2)$ |
| $\mathrm{Err}_h(\boldsymbol{\theta}_3)$ | $\mathrm{Err}_g(\boldsymbol{\gamma}_3)$ |
| $\mathrm{Err}_h(\boldsymbol{\theta}_4)$ | $\mathrm{Err}_g(\boldsymbol{\gamma}_4)$ |
| $\mathrm{Err}_h(\boldsymbol{\theta}_5)$ | $\mathrm{Err}_g(\boldsymbol{\gamma}_5)$ |

Model 1 is better iff.

$$\frac{1}{K}\sum_i Err_h(\theta_i) < \frac{1}{K}\sum_i Err_g(\gamma_i)$$

Hint: Go through homework problems 2.6, 2.7, 2.10, 3.3, 3.4 and similar problems in Hmw 4