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

Introduction to Data Science

Lecture 21 “Supervised” Learning

Zicheng Wang

Recap

How Does Machine Learning Work?

Step 1: Choose and Prepare a Training Data Set

- Training data consists of representative samples that a machine learning application uses to tune its model parameters.

Step 2: Select and Apply an Algorithm to the Training Data Set

- The type of machine learning algorithm you choose will primarily depend on the nature of the problem the model seeks to solve

Step 3: Model Training and Parameter Tuning

- Training the model involves adjusting the model's variables and parameters to enhance its accuracy in prediction.
- Training model does not require human intervention, showcasing the power of machine learning. The machine learns from the data, needing minimal to no guidance from the user.

Step 4: Deployment and Model Improvement

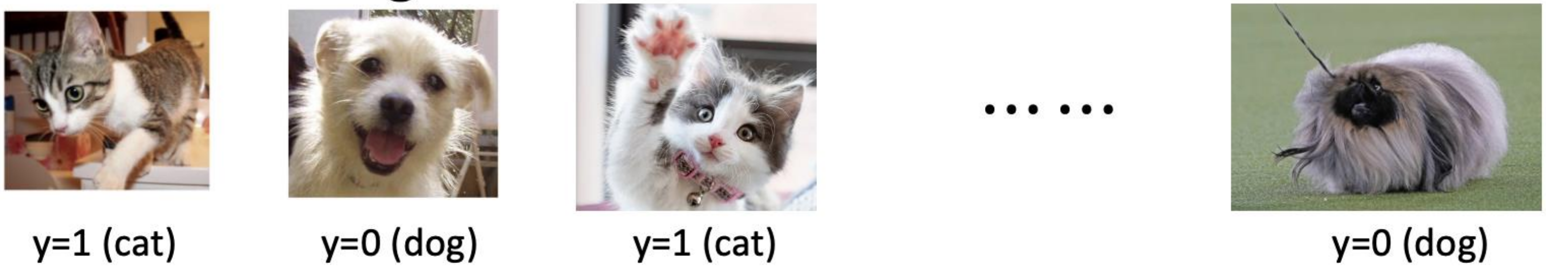
- Now you can deploy the mode for actual use and improve its effectiveness and accuracy over time with new data.

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, $(x^i, y^i), i = 1, \dots N$, you can observe both the input data x^i and the label y^i

Training data



Supervised Learning

Training data

y=1 (cat)

y=0 (dog)

y=1 (cat)

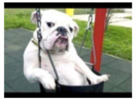
……

y=0 (dog)

⇩ Learning algorithm (optimization involved)

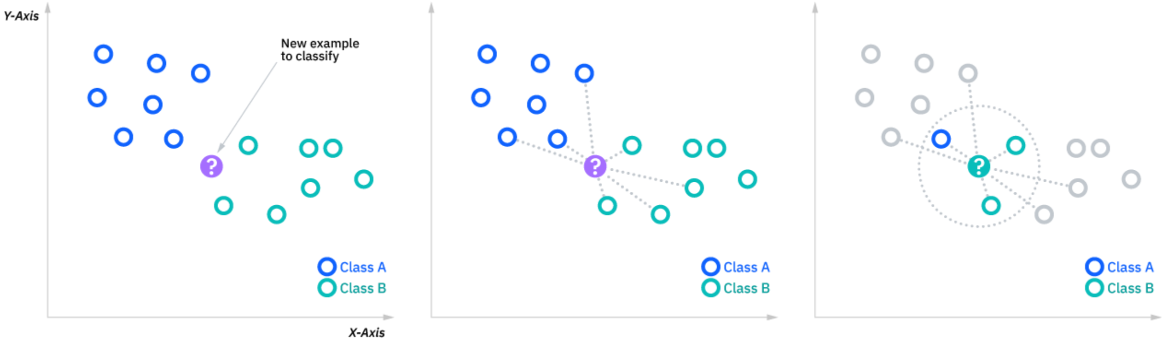
Classifier $h: X \rightarrow \{0,1\}$

For example:

$h\Big($  $\Big)$ ⟶ 0

K-Nearest Neighbor Classifier

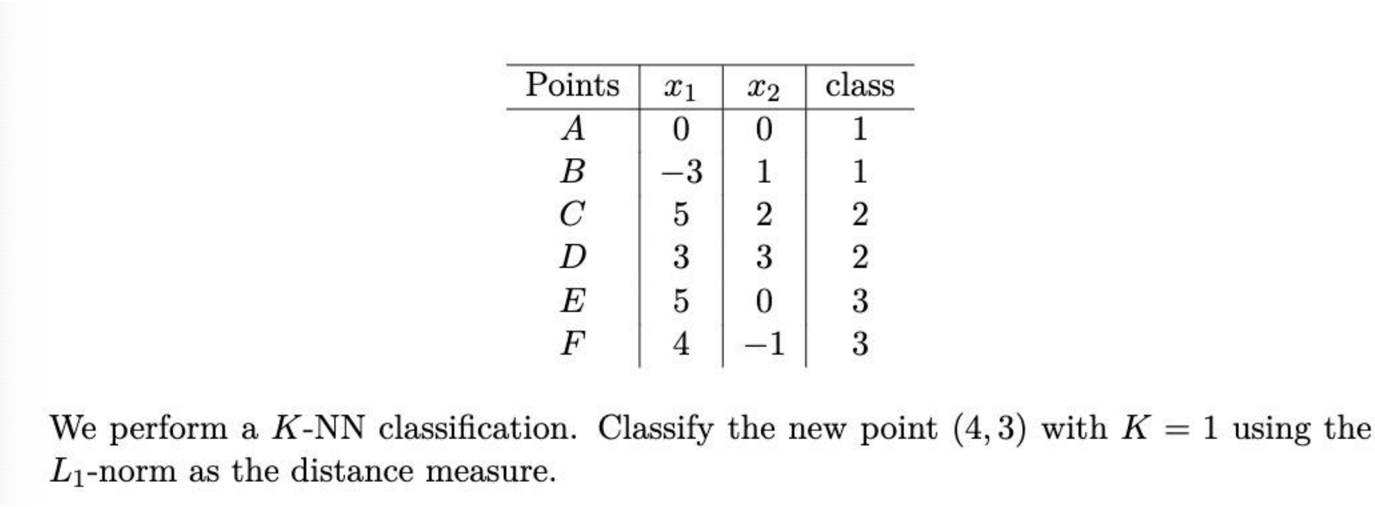
- Find K training points x_i closest to x.
- If the majority of K-nearest neighbors of x belong to classifier c, label x as c.



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

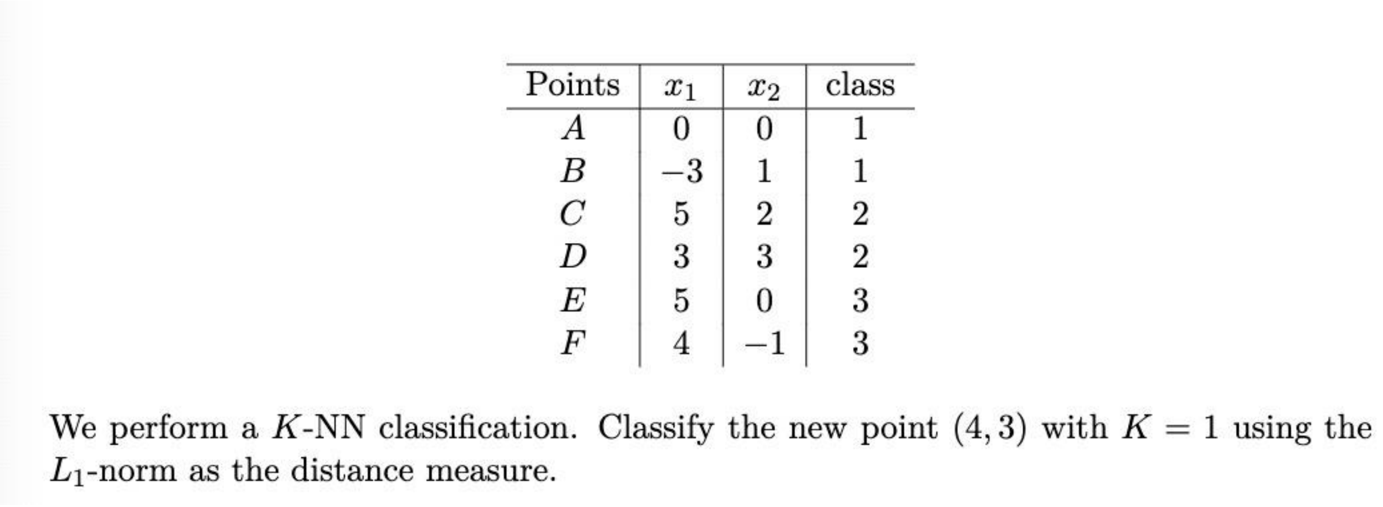
- We are given the following data set with points of three different classes:



Points	x_1	x_2	class
A	0	0	1
B	−3	1	1
C	5	2	2
D	3	3	2
E	5	0	3
F	4	−1	3

We perform a K-NN classification. Classify the new point $(4, 3)$ with $K = 1$ using the L_1-norm as the distance measure.

Manhattan distance between x and y: $|x_1 - y_1| + |x_2 - y_2|$



Points	x_1	x_2	class
A	0	0	1
B	-3	1	1
C	5	2	2
D	3	3	2
E	5	0	3
F	4	-1	3

We perform a K-NN classification. Classify the new point $(4,3)$ with $K=1$ using the L_1-norm as the distance measure.

$$\mathrm{d}(A, new\ point) = |0-4| + |0-3| = 7$$

$$\mathrm{d}(B, new\ point) = |-3-4| + |1-3| = 9$$

$$\mathrm{d}(C, new\ point) = |5-4| + |2-3| = 2$$

$$\mathrm{d}(D, new\ point) = |3-4| + |3-3| = 1$$

$$\mathrm{d}(E, new\ point) = |5-4| + |0-3| = 4$$

$$\mathrm{d}(F, new\ point) = |4-4| + |-1-3| = 4$$

Logistic Regression

Supervised Learning

Training data

y=1 (cat)

y=0 (dog)

y=1 (cat)

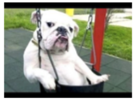
…… …

y=0 (dog)

Learning algorithm (optimization involved)

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

For example:

$h\Bigg($  $\Bigg)$ ⟶ $\boxed{0}$

Instead of letting h output {0,1}, we can consider that given x, h outputs the probability of each class

What is logistic regression model?

- Model the conditional probability of the label given the data

$P(it\ is\ a\ dog\ |$ $) = ?$

- Use all labeled samples to estimate the parameters of the conditional probability model.

What is logistic regression model?

- **Simplest** case (two classes): y ∈ $\{0\ ,1\}$

- **Logistic regression model:**

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

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

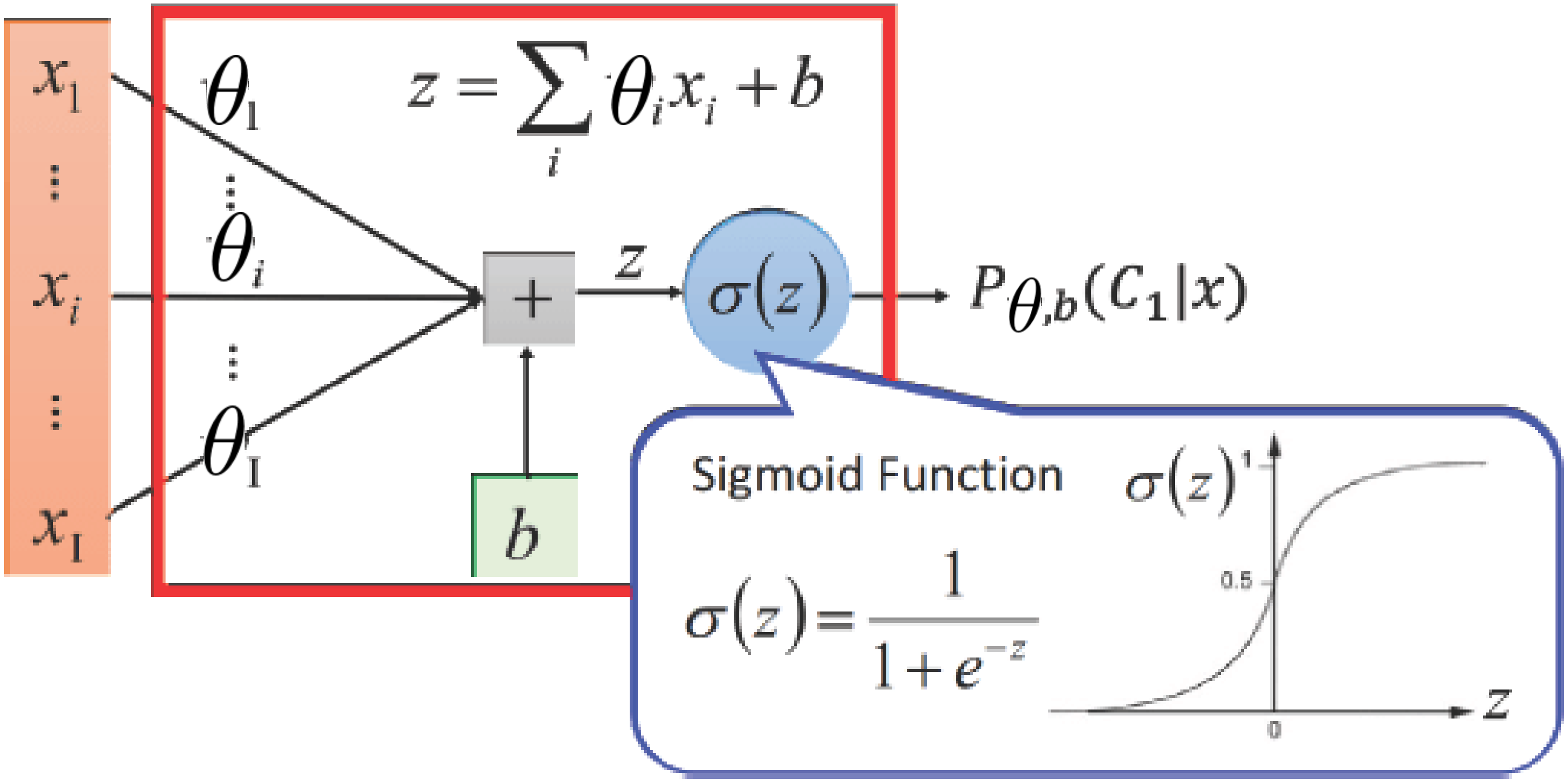
What is logistic regression model?

- **Simplest** case (two classes): y ∈ $\{0, 1\}$

- **Logistic regression model:**

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

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



For one-dimensional x, $p(y = 1|x, \theta, b) = \frac{1}{1+\exp(-(\theta x+b))}$

Train the Model

- **How to find $\boldsymbol{\theta}$ *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

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

Good news: $l(\boldsymbol{\theta}, b)$ is concave in $(\boldsymbol{\theta}, b)$

a single global optimum

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

- $\log P\left(y^i \middle| \boldsymbol{x}^i, \boldsymbol{\theta}, b\right)$ =$\left(y^i - 1\right)(\boldsymbol{\theta}^\top \boldsymbol{x}^i + b) - \log\left(1 + \exp(-\boldsymbol{\theta}^\top \boldsymbol{x}^i - b)\right)$

- Prove that $-\log\left(1 + \exp(-\boldsymbol{\theta}^\top \boldsymbol{x} - b)\right)$ is concave or $\log\left(1 + \exp(-\boldsymbol{\theta}^\top \boldsymbol{x} - b)\right)$ is convex.

- If the above holds, $l(\boldsymbol{\theta}, b) := \sum_{i=1}^{m} \log P\left(y^i \middle| \boldsymbol{x}^i, \boldsymbol{\theta}, b\right)$ is concave

Good news: $l(\boldsymbol{\theta}, b)$ is concave in $(\boldsymbol{\theta}, b)$

a single global optimum

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

- $\log P\left(y^i \middle| \boldsymbol{x}^i, \boldsymbol{\theta}, b\right) = \left(y^i - 1\right)(\boldsymbol{\theta}^\top \boldsymbol{x}^i + b) - \log\left(1 + \exp(-\boldsymbol{\theta}^\top \boldsymbol{x}^i - b)\right)$

- Prove that $-\log\left(1 + \exp(-\boldsymbol{\theta}^\top \boldsymbol{x} - b)\right)$ is concave or $\log\left(1 + \exp(-\boldsymbol{\theta}^\top \boldsymbol{x} - b)\right)$ is convex.

If $f_1, f_2, \ldots, f_n$ are convex functions, then $f_1 + f_2 + \cdots + f_n$ is also convex

- If the above holds, $l(\boldsymbol{\theta}, b) := \sum_{i=1}^{m} \log P\left(y^i \middle| \boldsymbol{x}^i, \boldsymbol{\theta}, b\right)$ is concave

Proof: $\log\left(1+\exp(-\boldsymbol{\theta}^{\top}\boldsymbol{x}-b)\right)$ is convex in $(\boldsymbol{\theta}, b)$

- Consider a point $(\boldsymbol{\theta}_0, b_0)$ and a direction vector **e = (**$\boldsymbol{\theta}_1$,b$_1$)
- h(t) := $\log\{1+\exp[(-\boldsymbol{\theta}_0^T\boldsymbol{x}-b_0)+t(-\boldsymbol{\theta}_1^T\boldsymbol{x}-b_1)]\}$

  = $\log\{1+\exp[C_1+tC_2]\}$
- **h''(t) =** $\dfrac{\boldsymbol{C_2^2}}{(\mathbf{1}+\exp[C_1+tC_2])^{\mathbf{2}}}\exp[C_1+tC_2]\geq\mathbf{0}$

Bad news: no closed form solution to the problem

$$l(\boldsymbol{\theta}, b) := \sum_{i=1}^{m} \left(y^i - 1\right)(\boldsymbol{\theta}^\top \boldsymbol{x}^i + b) - \log\left(1 + \exp(-\boldsymbol{\theta}^\top \boldsymbol{x}^i - b)\right)$$

- Gradient (θ is one-dimension)

$$\frac{\partial l(\theta,b)}{\partial \theta} = \sum_i \quad ((y^i - 1)\, x^i + \frac{\exp\left(-\theta x^i - b\right) x^i}{1 + \exp\left(-\theta x^i - b\right)})$$

$$\frac{\partial l(\theta,b)}{\partial b} = \sum_i \quad ((y^i - 1) + \frac{\exp\left(-\theta x^i - b\right)}{1 + \exp\left(-\theta x^i - b\right)})$$

- No closed form solution to the above system of equations

We need to use numerical methods to find $(\boldsymbol{\theta}^*, b^*)$ that maximizes $l(\boldsymbol{\theta}, b)$

Gradient Descent

Gradient Descent Method

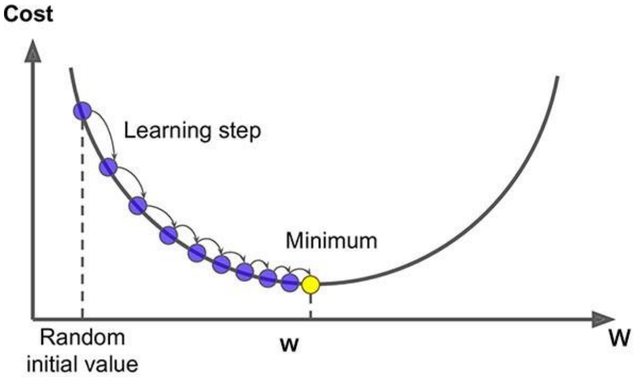
- Start with an initial point $x^{(0)}$

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

- Update our point by the following rule:

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

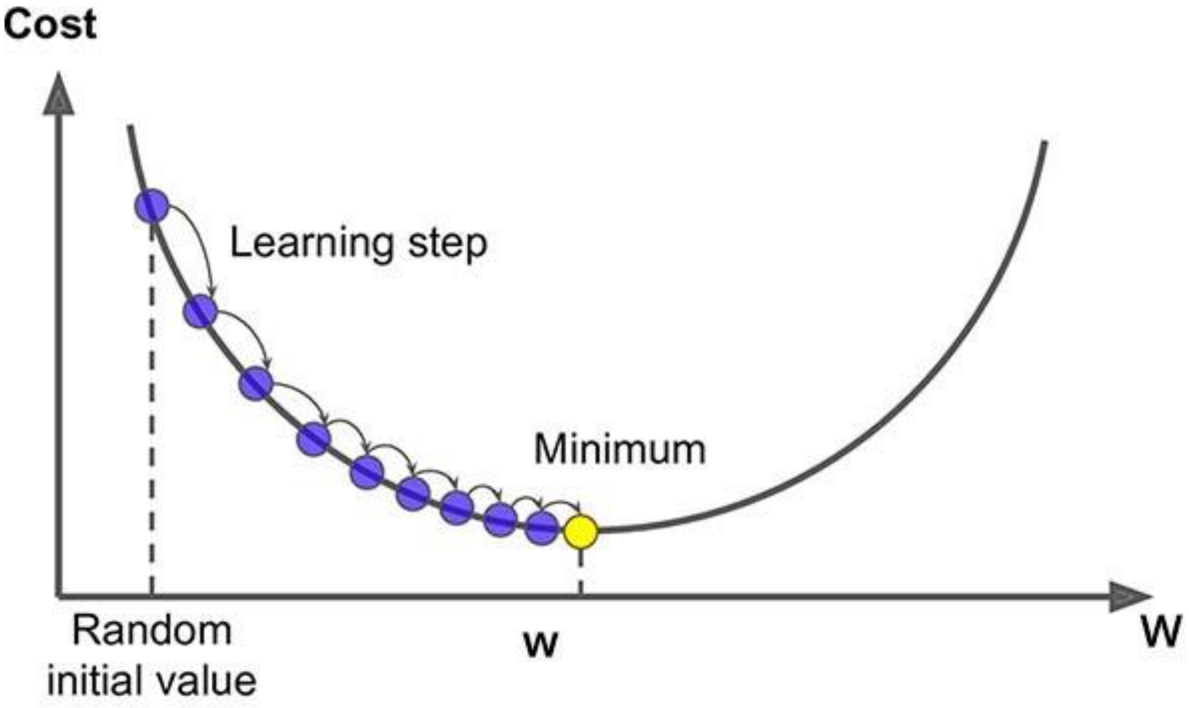
- Stopping criteria:
  - $|x^{(t+1)} - x^{(t)}| \leq \varepsilon$
  - or $|f'(x^{(t)})| \leq \varepsilon$



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.

Homer descending !

We may want the step size, $\alpha^{(t)}$, to be large during the initial steps and smaller as we approach the local minimizer



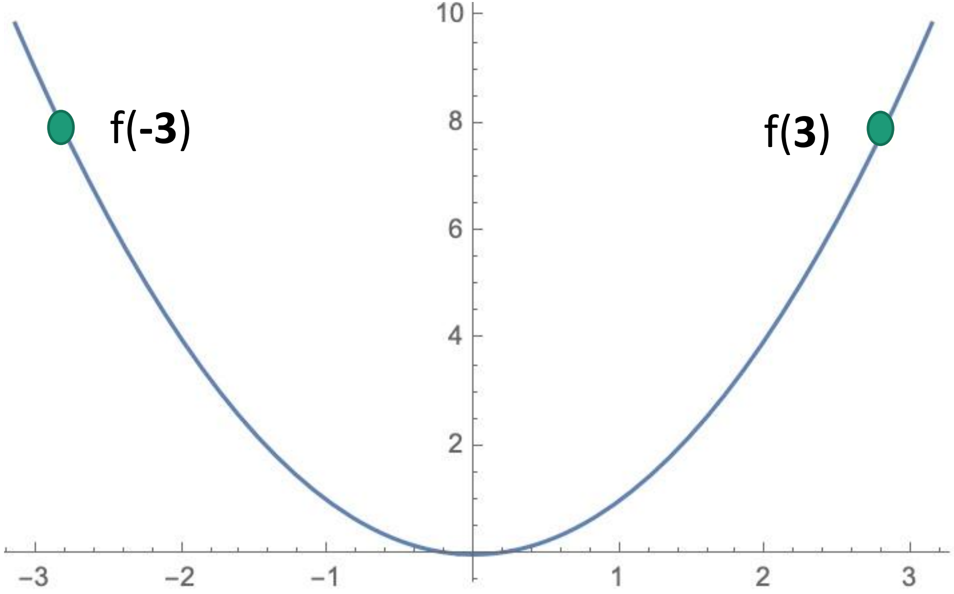
If $\boldsymbol{\alpha^{(t)}}$ is a constant, we may not meet the stop criteria.

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



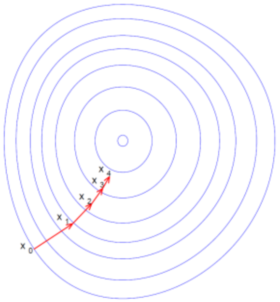
The updated points will oscillate between 3 and -3

Gradient Descent algorithm for logistic regression

- Initialize parameter $(\boldsymbol{\theta}^0, b^0)$
- While $|\theta^{t+1} - \theta^t| > \epsilon$ or $|b^{t+1} - b^t| > \epsilon$, Do

$$\boldsymbol{\theta}^{t+1} \leftarrow \boldsymbol{\theta}^t + \alpha^{(t)} \frac{1}{m} \sum_i (y^i - 1)\, \boldsymbol{x}^i + \frac{\exp\left(-{\boldsymbol{\theta}^t}^T \boldsymbol{x}^i - b^t\right) \boldsymbol{x}^i}{1 + \exp\left(-{\boldsymbol{\theta}^t}^T \boldsymbol{x}^i - b^t\right)}$$

$$b^{t+1} \leftarrow b^t + \alpha^{(t)} \frac{1}{m} \sum_i (y^i - 1) + \frac{\exp\left(-{\boldsymbol{\theta}^t}^T \boldsymbol{x}^i - b^t\right)}{1 + \exp\left(-{\boldsymbol{\theta}^t}^T \boldsymbol{x}^i - b^t\right)}$$



A variant: Stochastic gradient descent

- At each iteration, we randomly choose a small batch of samples in the training data set, and update using the stochastic gradient

$$\boldsymbol{\theta}^{t+1} \leftarrow \boldsymbol{\theta}^{t} + \alpha^{(t)} \frac{1}{|B|} \sum_{i \in B} \left((y^i - 1)\, \boldsymbol{x}^i + \frac{\exp\left(-{\boldsymbol{\theta}^{t}}^{T} \boldsymbol{x}^i - b^t\right) \boldsymbol{x}^i}{1 + \exp\left(-{\boldsymbol{\theta}^{t}}^{T} \boldsymbol{x}^i - b^t\right)} \right)$$

$$b^{t+1} \leftarrow b^{t} + \alpha^{(t)} \frac{1}{|B|} \sum_{i \in B} \left((y^i - 1) + \frac{\exp\left(-{\boldsymbol{\theta}^{t}}^{T} \boldsymbol{x}^i - b^t\right)}{1 + \exp\left(-{\boldsymbol{\theta}^{t}}^{T} \boldsymbol{x}^i - b^t\right)} \right)$$

- B: the batch we use in each iteration

Summary

- KNN:
  - Use K nearest neighbors to determine the label.
  - We do not need to use all samples.

- Logistic regression:
  1. Propose a logistic model (conditional probability)
  2. Use all samples to estimate the parameters
  3. Use the estimated model to calculate the conditional probability for new data
  4. Numerical methods to find the solution (Gradient descent)