Lecture 7: Logistic Regression
Winter 2018

Kai-Wei Chang

CS @ UCLA

kw+cm146@kwchang.net

The instructor gratefully acknowledges Dan Roth, Vivek Srikuar, Sriram Sankararaman, Fei Sha, Ameet Talwalkar, Eric Eaton, and Jessica Wu whose slides are heavily used, and the many others who made their course material freely available online.

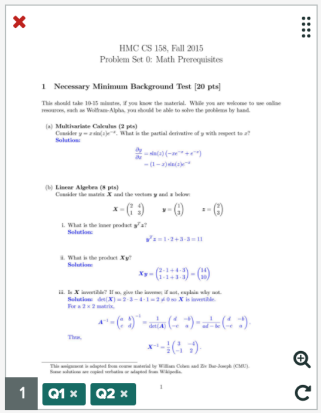
Announcement

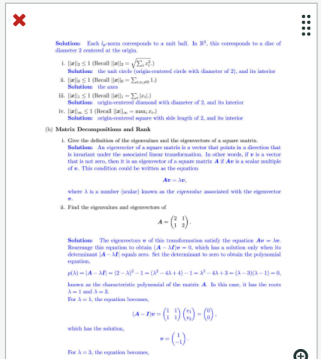
- Problem set 1 due today
- Please help to locate your answers

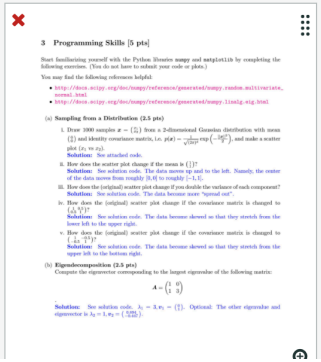
Question Outline

Select pages to assign to Question **5**.

TITLE	POINTS
1 Calculus P1 ×	2.0 pts
2 Linear Algebra P1 × P2 ×	8.0 pts
3 Probability P2 ×	10.0 pts
4 probability axioms P3 × P4 ×	5.0 pts
5 Distributions	5.0 pts
6 Mean and variance	5.0 pts
7 Algorithm	10.0 pts
8 Random variables	5.0 pts
9 Linear Algebra	20.0 pts
10 sampling from a distribution	2.5 pts
11 Eigen decomposition	2.5 pts
12 Data	5.0 pts

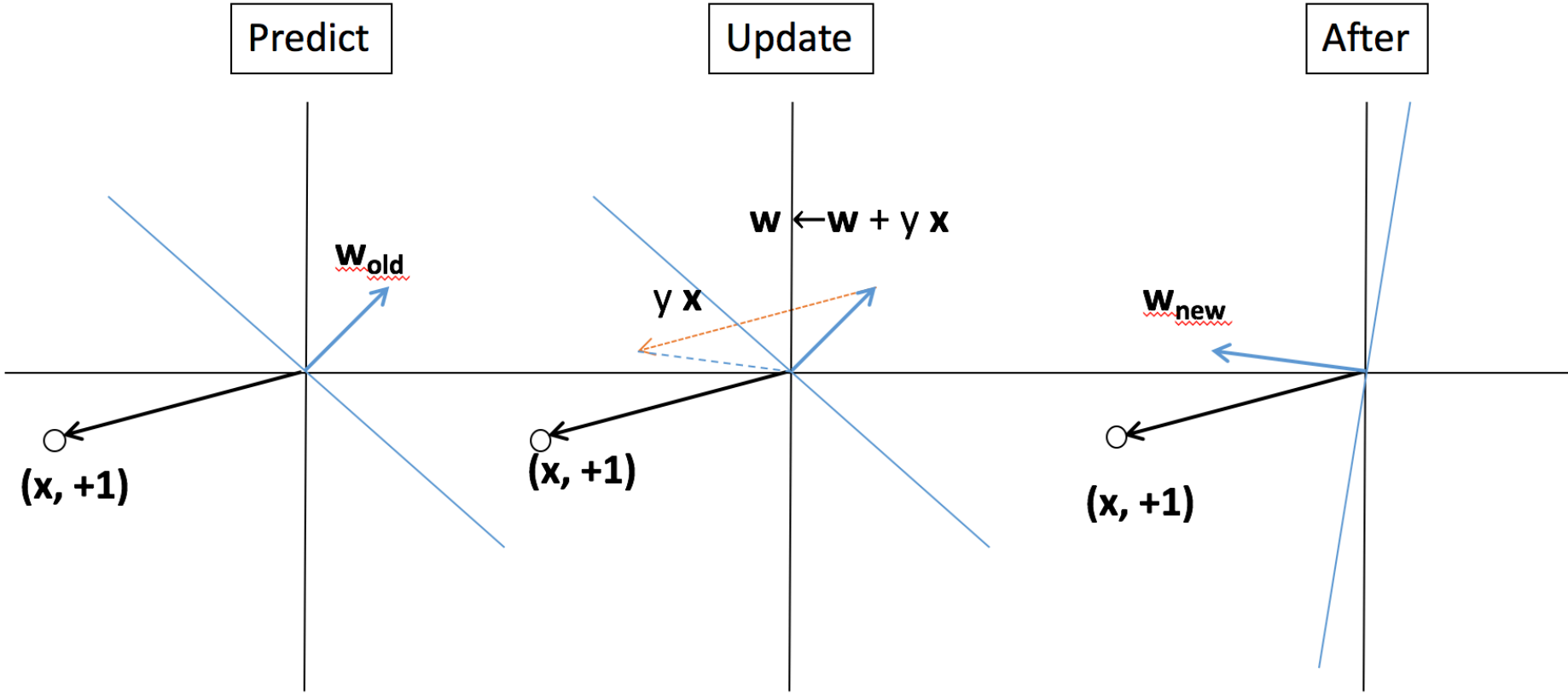




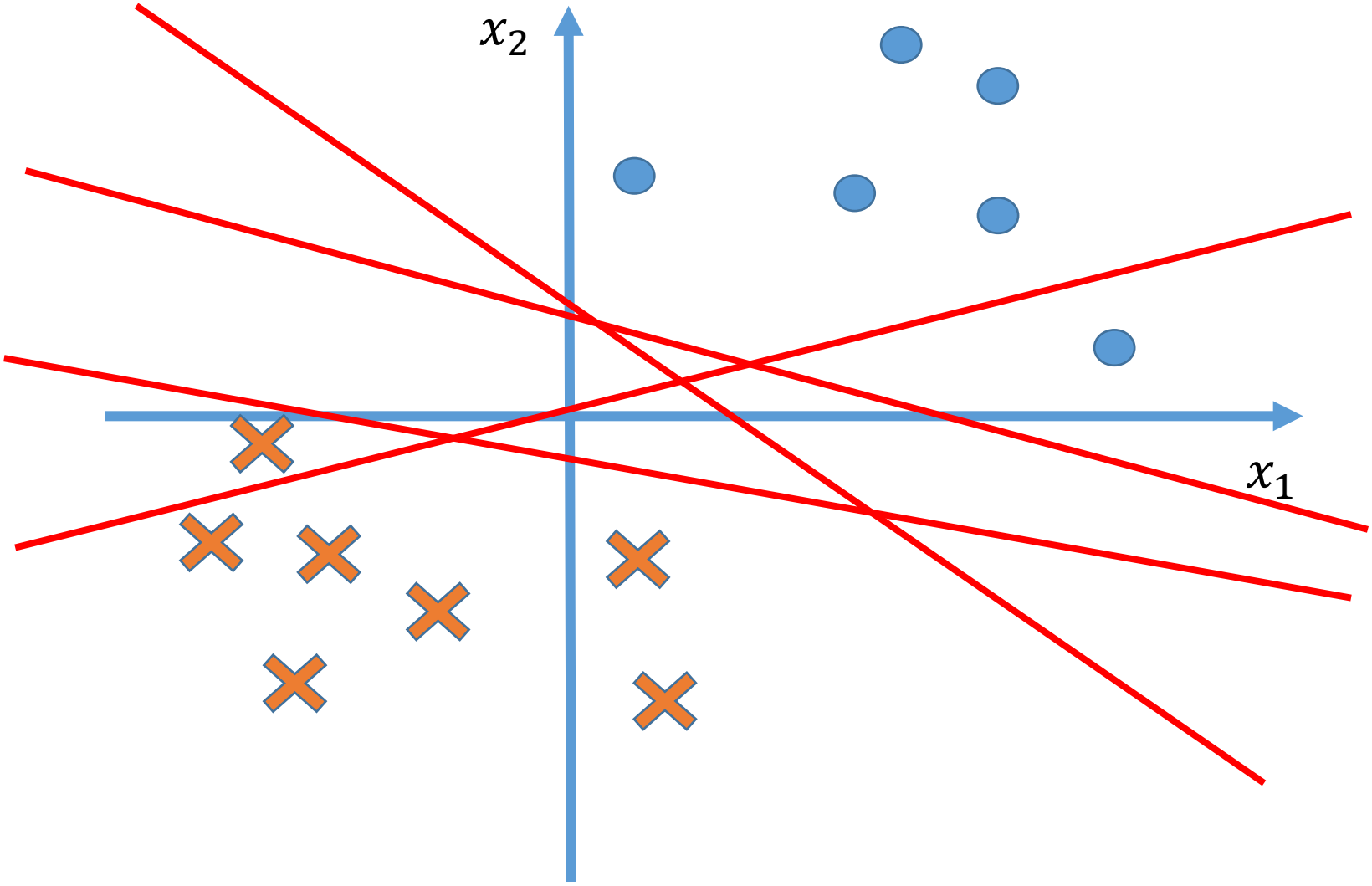


Recap: Perceptron

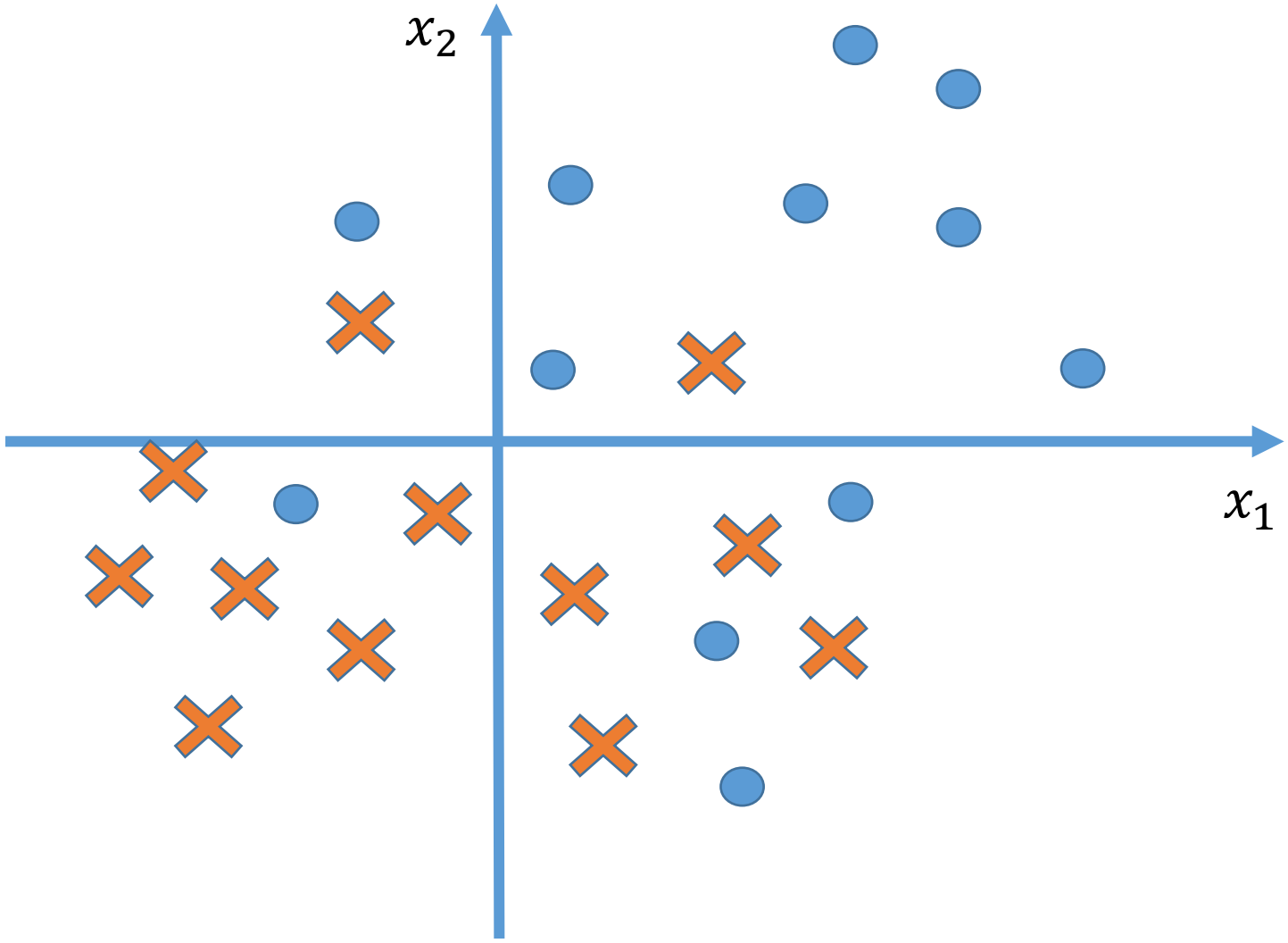
- Learning a linear separator when data is linearly separable.



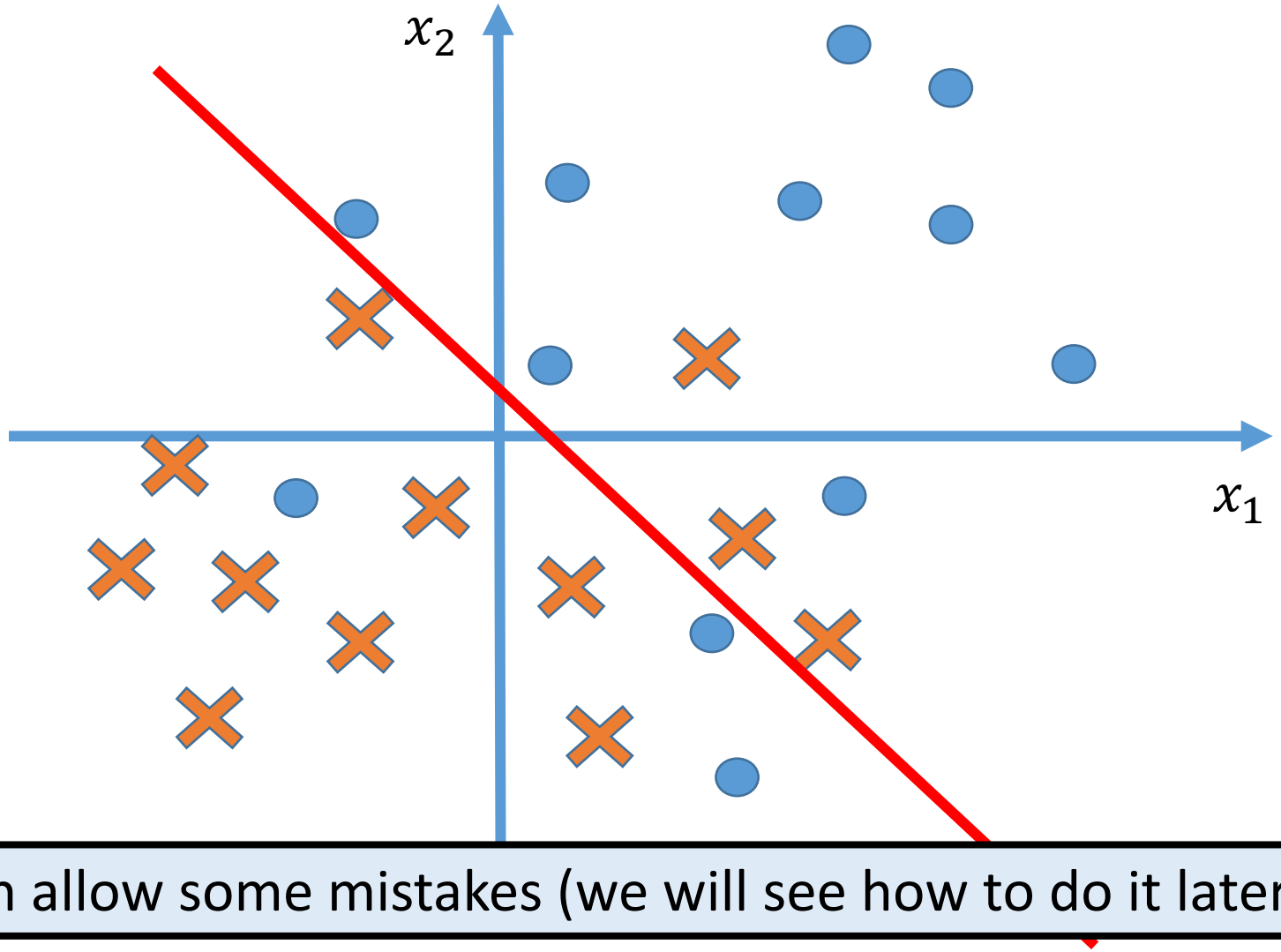
Hypothesis space: linear model



What if data is not linearly separable?

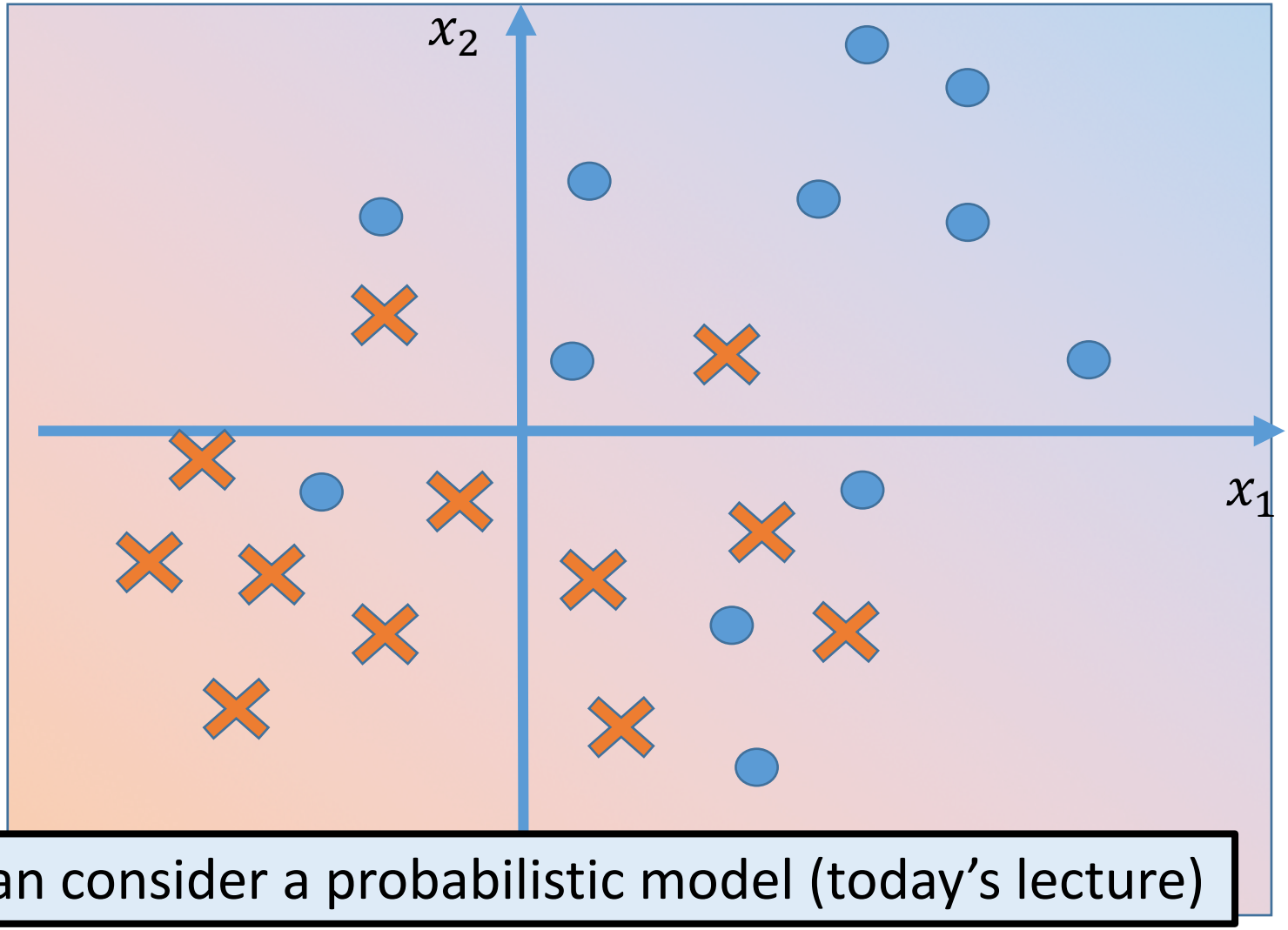


What if data is not linearly separable?



We can allow some mistakes (we will see how to do it later)

What if data is not linearly separable?



We can consider a probabilistic model (today's lecture)

What you will learn today

- Logistic Regression
  - How to model – hypothesis space, model definition
  - How to predict?
  - How to train?
- Maximum Likelihood
- Gradient descent

Logistic Regression: Setup

- The setting
  - Binary classification
  - Inputs: Feature vectors $x \in R^N$
  - Labels: $\mathbf{y} \in \{-1, +1\}$
- Training data
  - $S = \{(\mathbf{x}_i, \mathbf{y}_i)\}$, m examples
- Goal
  - Linear model

Goal: Learn a Linear Classifiers

- *Linear Threshold Units* (LTUs) classify an example **x** using the following classification rule
  - Output = $\text{sgn}(\mathbf{w}^T\mathbf{x} + b) = \text{sgn}(b + \sum w_i x_i)$
  - $\mathbf{w}^T\mathbf{x} + b \geq 0$ Predict y = 1
  - $\mathbf{w}^T\mathbf{x} + b < 0$ Predict y = -1

Tie is broken arbitary

b is called the bias term

Recap: A trick to remove the bias term b

$$
\begin{aligned}
& w^T x + b \\
&= [w^T \; b] \cdot \begin{bmatrix} x \\ 1 \end{bmatrix} \\
&= \tilde{w} \cdot \tilde{x}
\end{aligned}
$$

$$
\begin{aligned}
\tilde{w} &= [w^T \; b]^T \\
\tilde{x} &= [x^T \; 1]^T
\end{aligned}
$$

For simplicity, I may write $\tilde{w}$ and $\tilde{x}$ as w and x when there is no confusion

What is our hypothesis space?

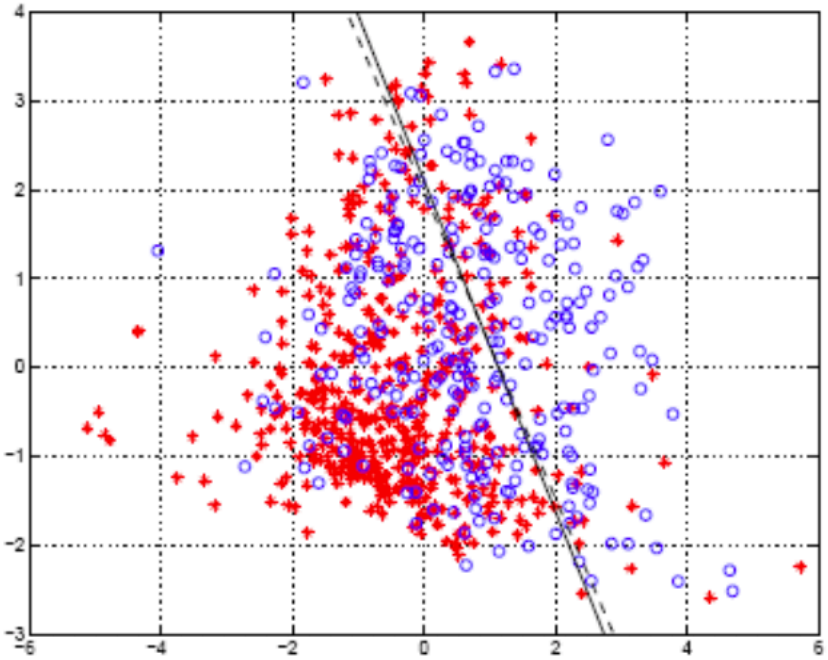
- Formally, we are consider all functions in the following space:

$$H = \{\ \mathrm{h} \mid h : X \to Y, h(x) = sgn(\ w^T x + b)\ \}$$

- Learning goal: $y = h(x)$
  - Learn w, b (w and x are n-dimensional vectors)
  - w: weight vector, b: bias

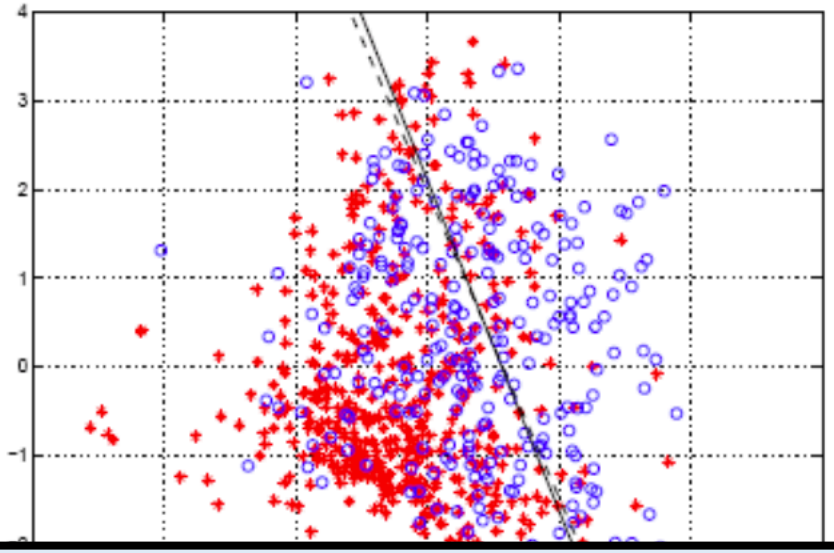
Classification, but…

- The output **y** is discrete valued (-1 or 1)
- Instead of predicting the output, let us try to predict P(y = 1 | **x**)



Classification, but…

- The output $\mathbf{y}$ is discrete valued (-1 or 1)
- Instead of predicting the output, let us try to predict P(y = 1 | **x**)



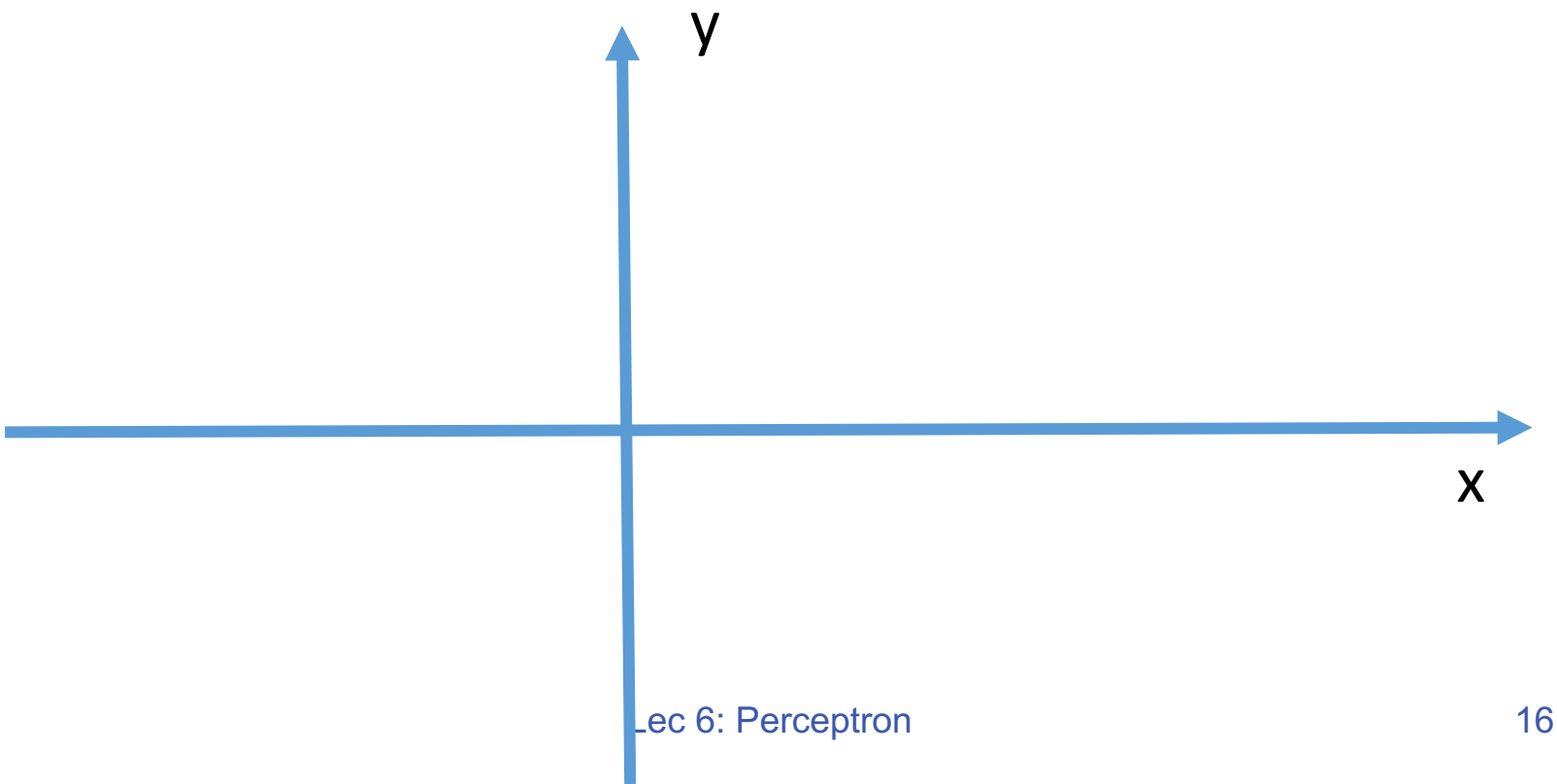
Perceptron does not produce probability estimates

Predict P(y = 1 | **x**)

- Expand hypothesis space to functions whose output is [0-1]
- Original problem: $h(x) \in \{-1, 1\}$
- Modified problem: $h(x) \in [0\text{-}1]$
- Effectively make the problem a regression problem

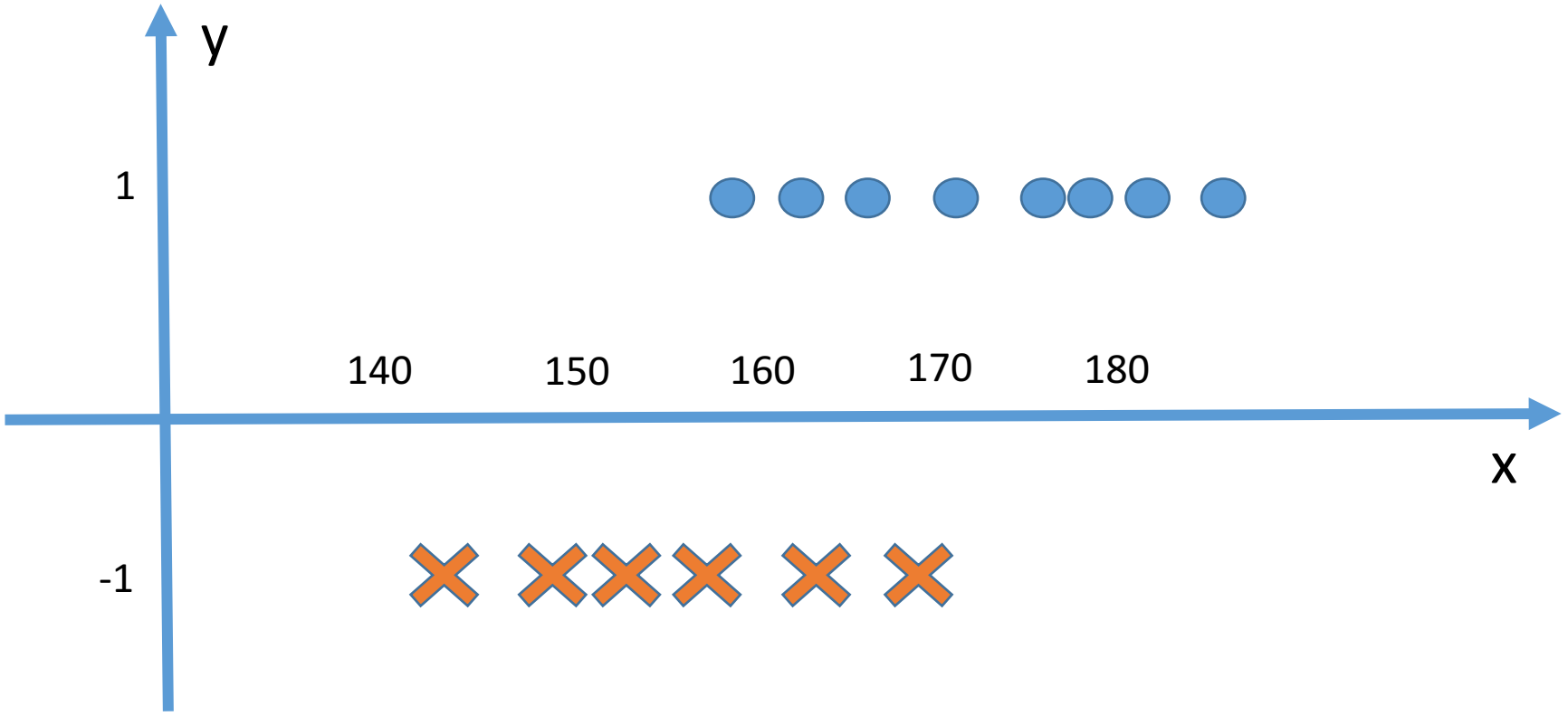
How to make the prediction within [0,1]

- What does data look like
- Assume we only have one feature (e.g., predicting gender by height



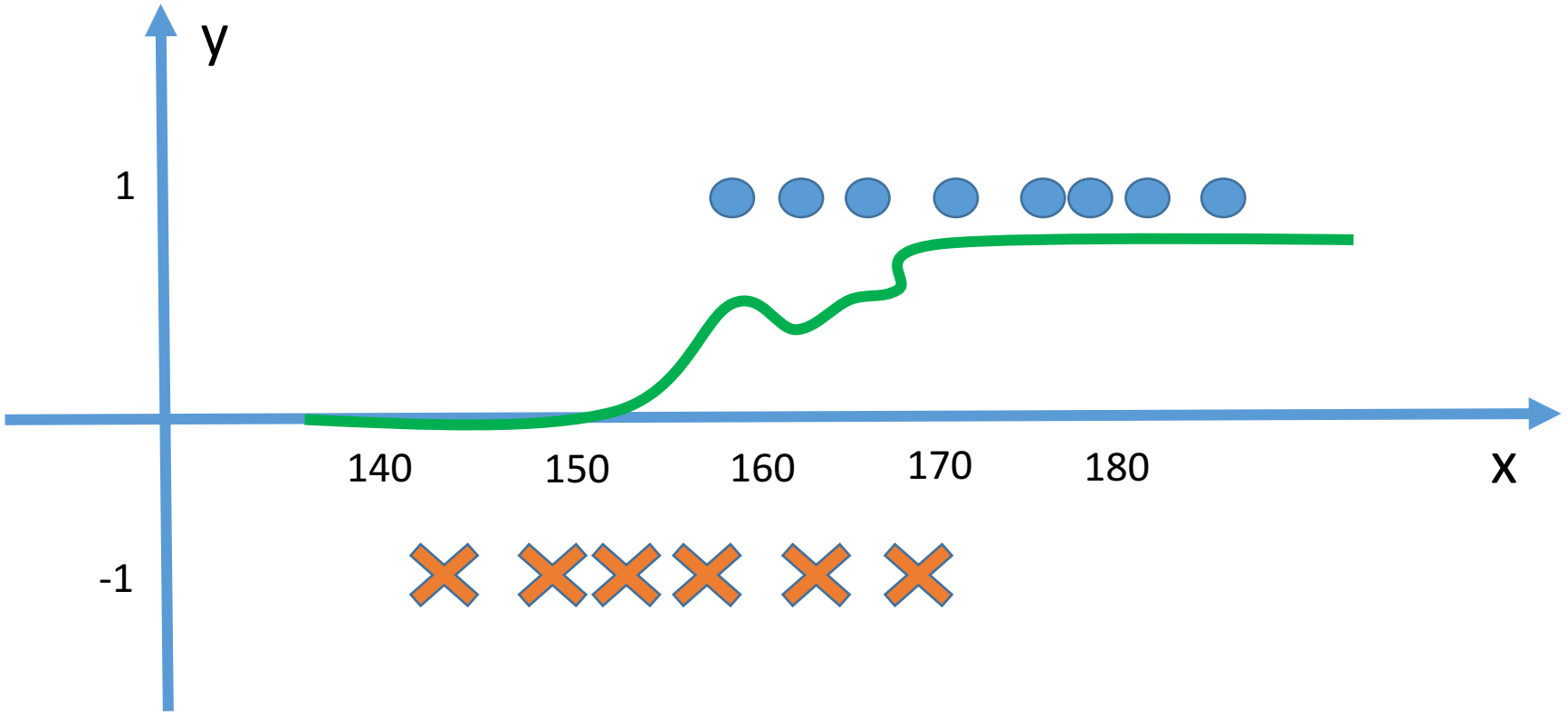
How to make the prediction within [0,1]

- What does data look like
- Assume we only have one feature



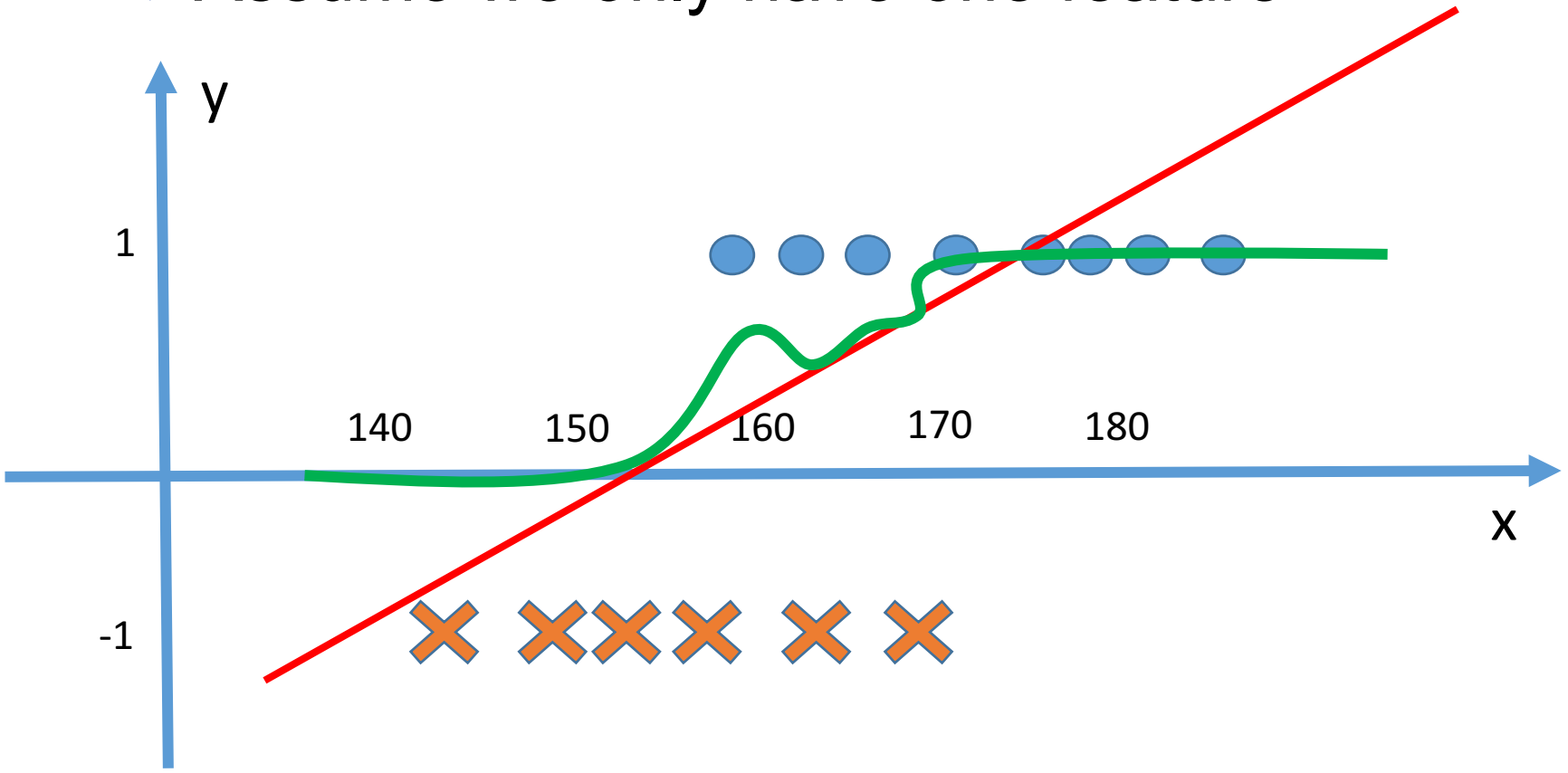
How to make the prediction within [0,1]

- What does $P(y = 1|x)$ look like
- Assume we only have one feature



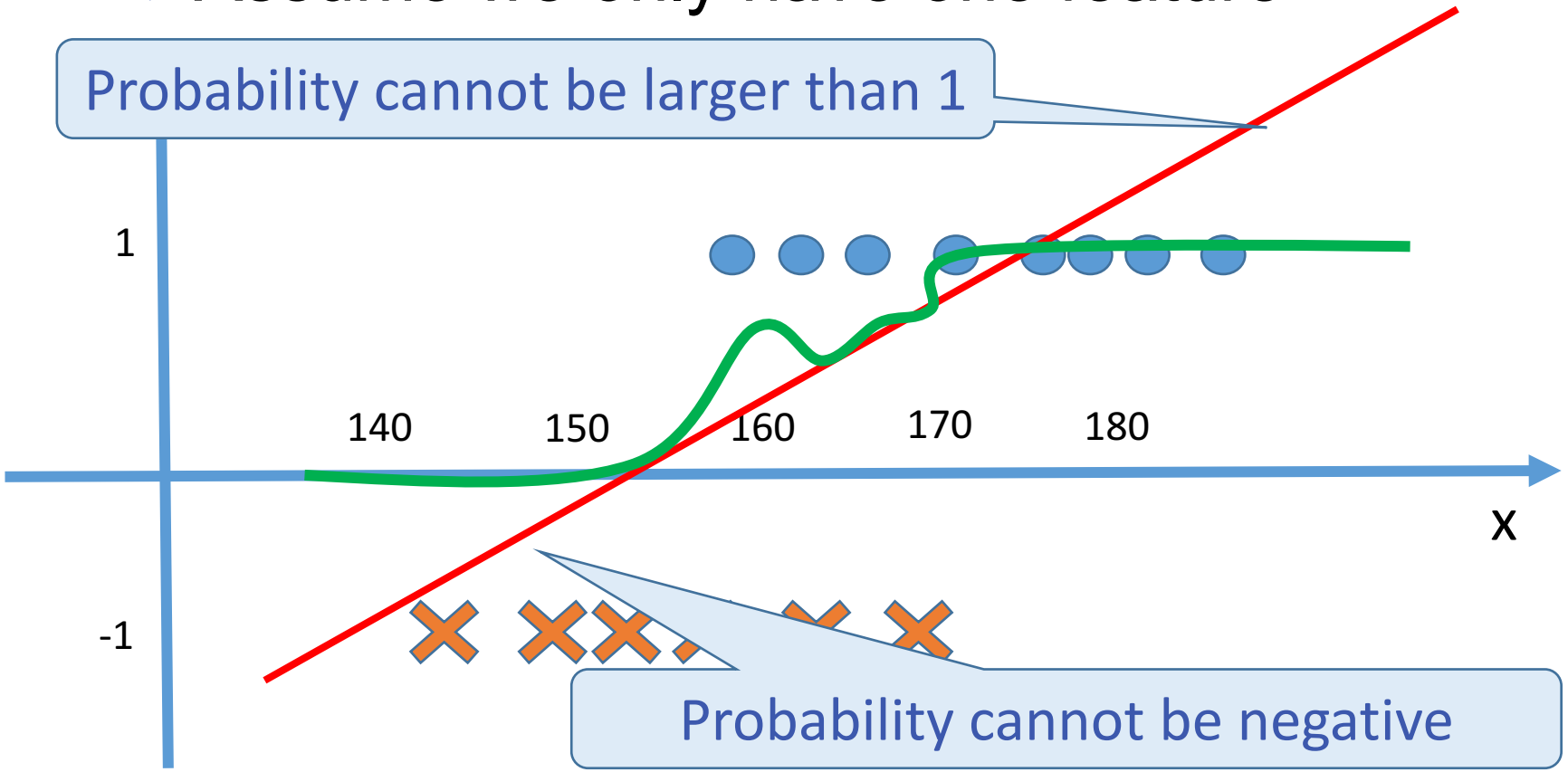
How to make the prediction within [0,1]

- What does $y = \mathbf{w}^T\mathbf{x} + b$ look like
- Assume we only have one feature



How to make the prediction within [0,1]

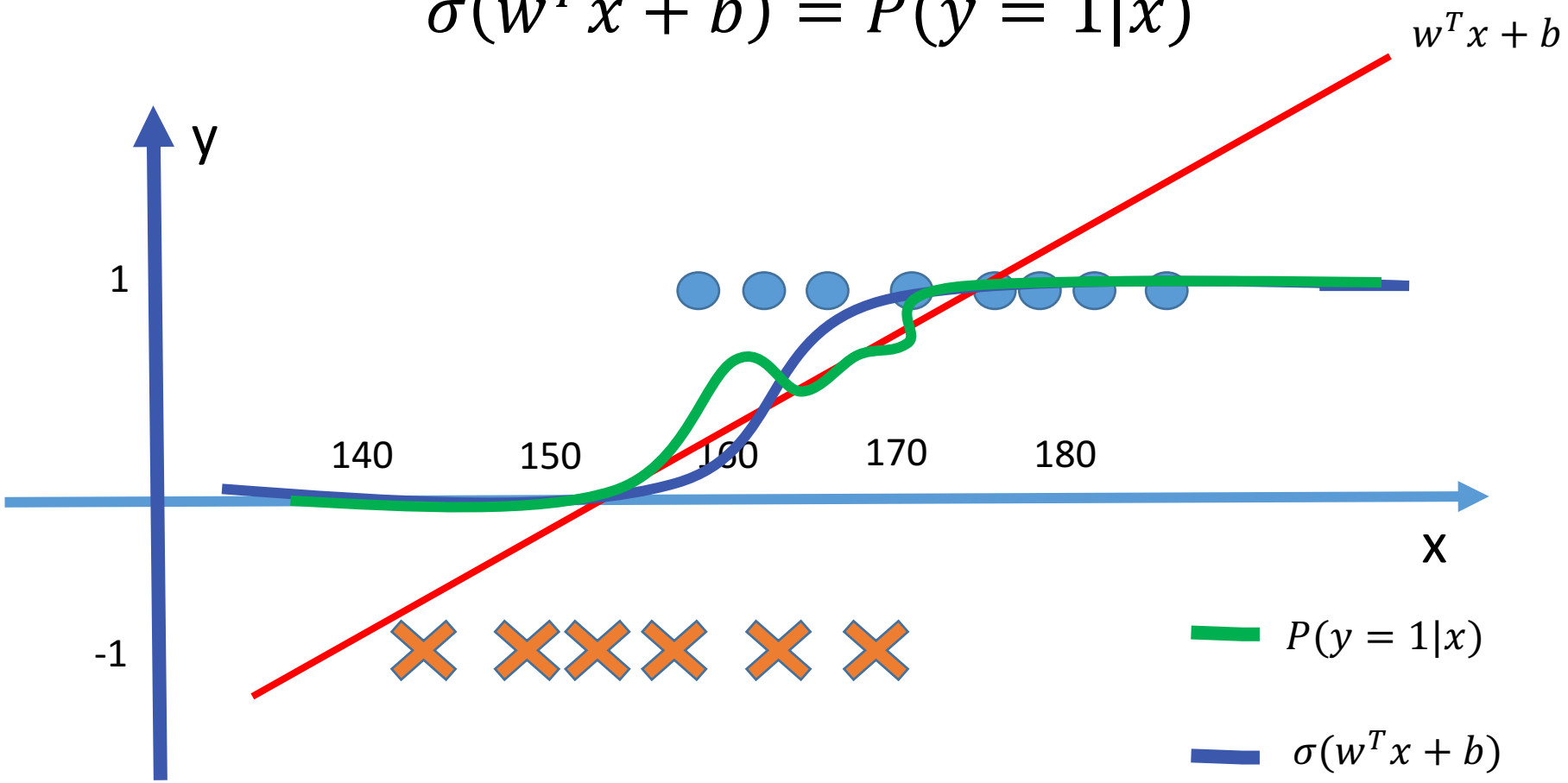
- What does $y = \mathbf{w}^T\mathbf{x} + b$ look like
- Assume we only have one feature



Define a linear model to fit the curve

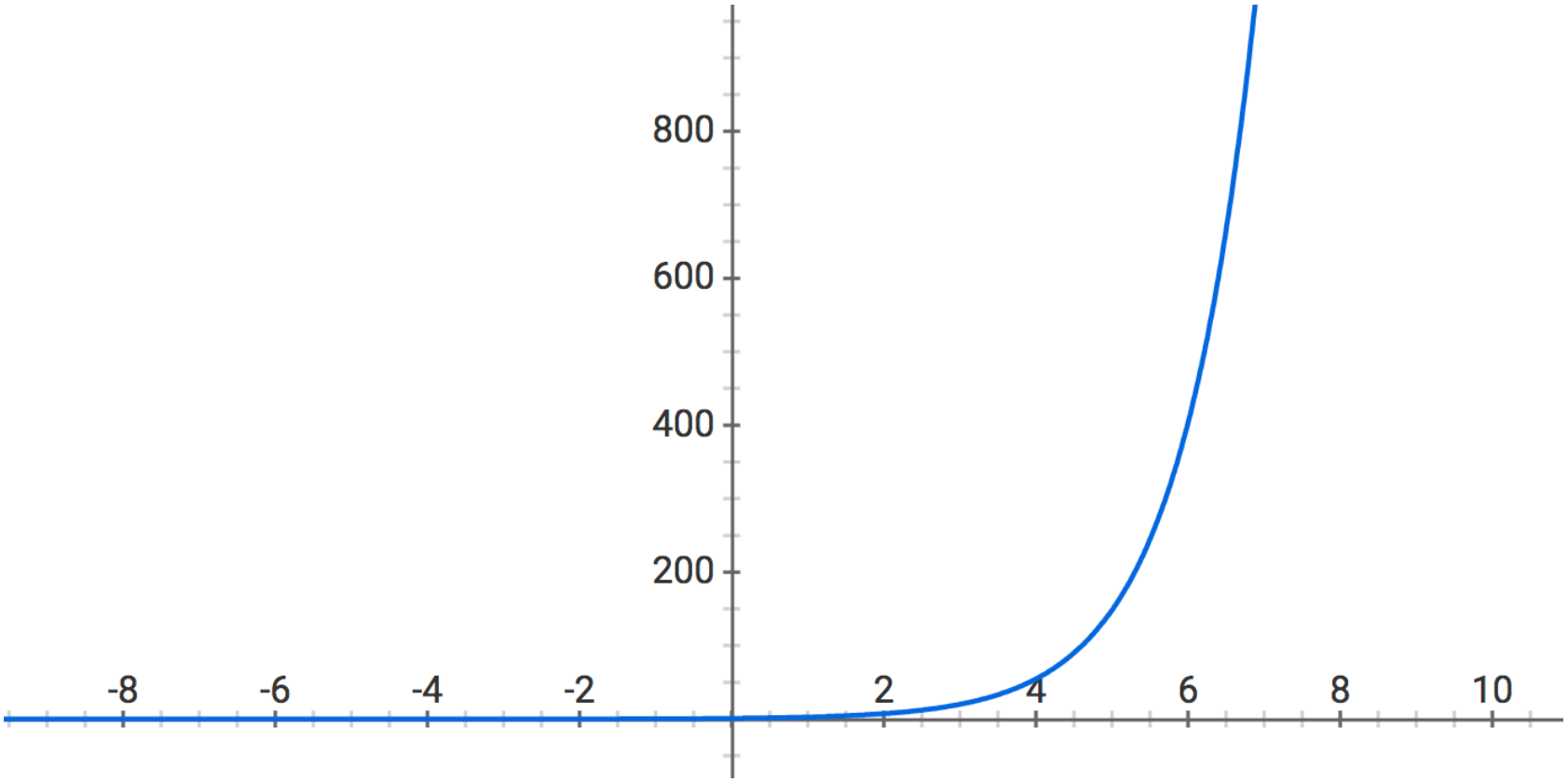
- Define a transformation function such that

$$\sigma(w^T x + b) = P(y = 1|x)$$



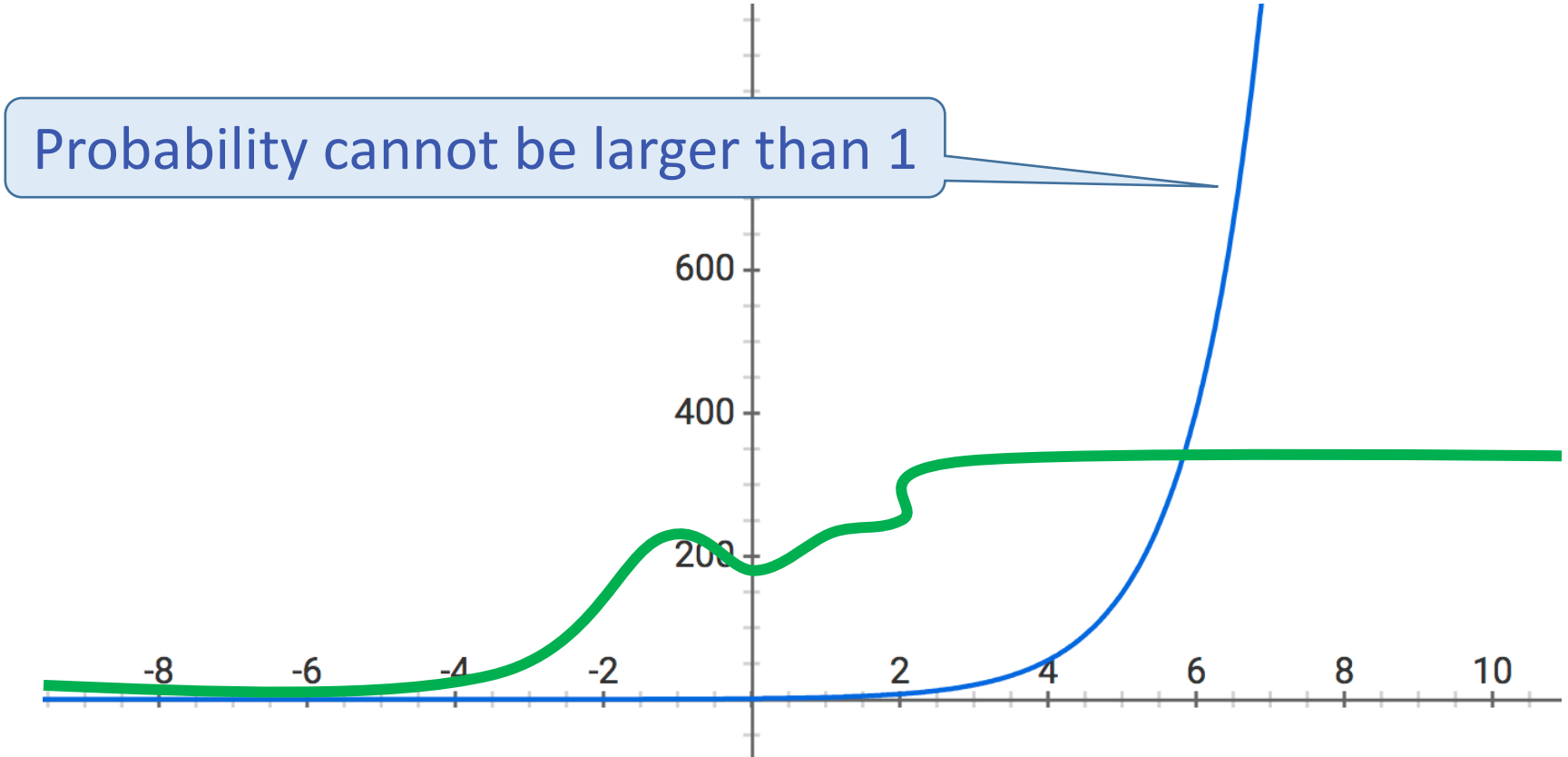
How can we design such a transformation function?

- Idea 1: function always output positive value
  - $\exp(\cdot)$ is always positive



How can we design such a transformation function?

- Idea 1: function always output positive value
  - $\exp(\cdot)$ is always positive
  - $\exp(w^T x + b)$ always return positive value



How can we design such a transformation function?

- Idea 2: normalize the value such that it is less than 1
  - $\exp(w^T x + b) \in (0, \infty)$ grows very fast, so we need to use exp to normalize itself
  - Let's use

$$\sigma(w^T x + b) = \frac{\exp(w^T x + b)}{1 + \exp(w^T x + b)}$$

  - When $w^T x + b \to \infty$,

$$\sigma(w^T x + b) \to 1$$

  - When $w^T x + b \to -\infty$,

$$\sigma(w^T x + b) \to 0$$

The Sigmoid function

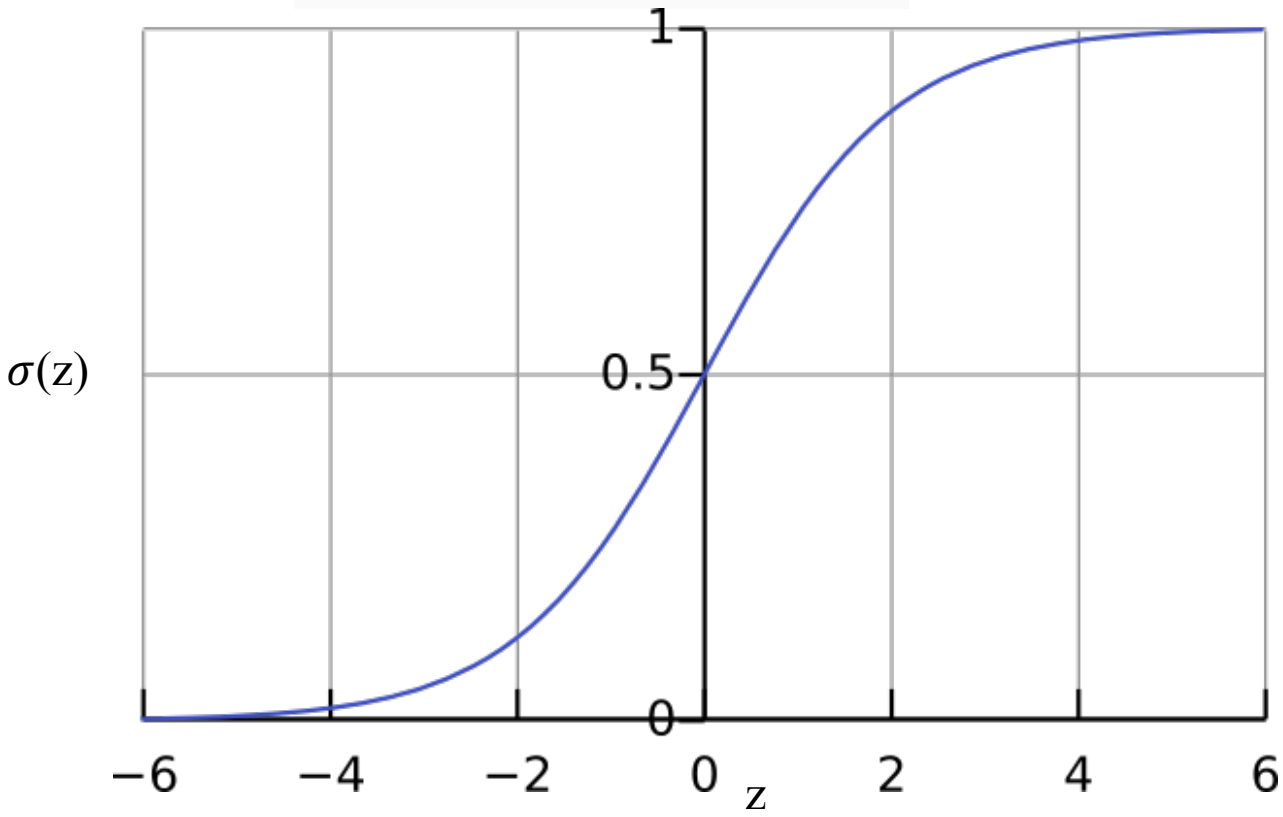
The $\sigma(z)$ function is called sigmoid function (or logistic function)

$$\sigma(z) = \frac{\exp(z)}{1 + \exp(z)}$$
$$= \frac{1}{1+\exp(-z)}$$

What is the domain and the range of the sigmoid function?

The Sigmoid function

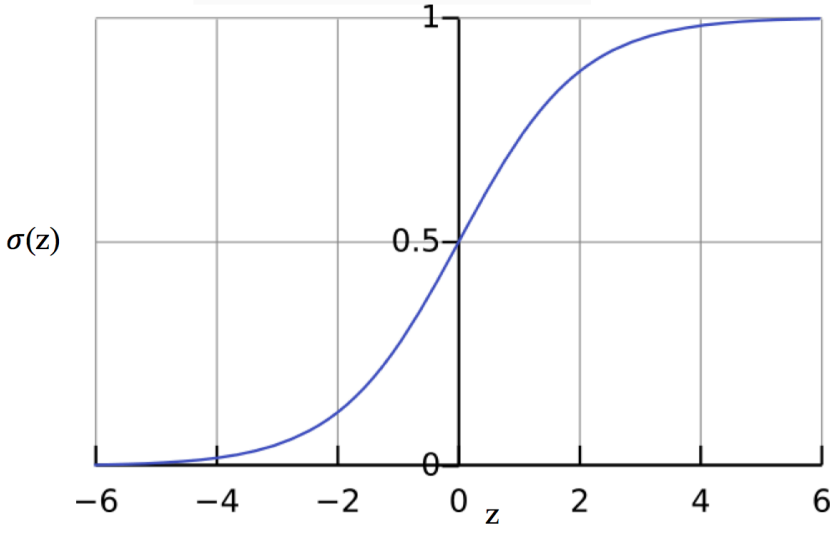
$$\sigma(z) = \frac{1}{1 + \exp(-z)}$$



logistic function or sigmoid function

- When $z \to \infty$ what is $\sigma(z)$?
- When $z \to -\infty$ what is $\sigma(z)$?
- When $z = 0$ what is $\sigma(z)$?

$$\sigma(z) = \frac{1}{1 + \exp(-z)}$$



The properties of Sigmoid function

$$\sigma(z) = \frac{1}{1 + \exp(-z)}$$

Exercise

What is its derivative with respect to z?

$$\frac{d\sigma}{dz} = \frac{d}{dz}\frac{1}{1 + \exp(-z)}$$

The Sigmoid function

$$\sigma(z) = \frac{1}{1 + \exp(-z)}$$

What is its derivative with respect to z?

$$\begin{aligned}
\frac{d\sigma}{dz} &= \frac{d}{dz}\frac{1}{1+\exp(-z)} \\
&= \frac{1}{(1+\exp(-z))^2} \cdot \exp(-z) \\
&= \left(1 - \frac{1}{1+\exp(-z)}\right) \cdot \frac{1}{1+\exp(-z)} \\
&= \sigma(z)\left(1 - \sigma(z)\right).
\end{aligned}$$

Nice computational properties

The hypothesis space for logistic regression

Let's ignore b from now by using the trick in page 11

All functions of the form

$$h_{\mathbf{w}}(\mathbf{x}) = \sigma(\mathbf{w}^T\mathbf{x}) = \frac{1}{1+\exp(-\mathbf{w}^T\mathbf{x})}$$

That is, a linear function, composed with a sigmoid function (the logistic function)

$$\sigma(z) = \frac{1}{1+\exp(-z)}$$

We want to find $h_w(x)$ such that

$$\mathrm{h}_w(x) \approx P(y=1\,|x)$$

Check point (Modeling)

- What is the goal of logistic regression?
  - Model $P(y = 1|x)$
- What is the hypothesis space?

$$H = \{\ \mathrm{h} \mid h : X \to P(Y \mid X), h(x) = \sigma\ (\ w^T x + b)\ \}$$

$$\sigma(z) = \frac{1}{1 + \exp(-z)}$$

Understanding a ML model

- What are the input / output
  - The input/output of data may be different from your model
  - E.g., feature selection may remove some features from data
- What is the hypothesis space?
- What is the loss function
  (criteria of model fitting)
- How to make prediction
- How to learn the model
- Properties / Guarantees of the model
- Pros/Cons/Connections to other models

Understanding a ML model

- What are the input / output
  - The input/output of data may be different from your model
  - E.g., feature selection may remove some features from data
- What is the hypothesis space?
- What is the loss function
  (criteria of model fitting)

modeling

- How to make prediction
- How to learn the model
- Properties / Guarantees of the model
- Pros/Cons/Connections to other models

Understanding a ML model

- What are the input / output
  - The input/output of data may be different from your model
  - E.g., feature selection may remove some features from data
- What is the hypothesis space?
- What is the loss function
  (criteria of model fitting)
- How to make prediction
- How to learn the model — algorithm
- Properties / Guarantees of the model
- Pros/Cons/Connections to other models

Understanding a ML model

- What are the input / output
  - The input/output of data may be different from your model
  - E.g., feature selection may remove some features from data
- What is the hypothesis space?
- What is the loss function
  (criteria of model fitting)
- How to make prediction
- How to learn the model

- Properties / Guarantees of the model
- Pros/Cons/Connections to other models

analysis

Logistic regression

- What are the input / output
- What is the hypothesis space?
- What is the loss function
  (criteria of model fitting)
- How to make prediction
- How to learn the model
- Properties / Guarantees of the model
- Pros/Cons/Connections to other models

Logistic regression

- What are the input / output
- What is the hypothesis space?
- What is the loss function
  (criteria of model fitting)
- How to make prediction
- How to learn the model
- Properties / Guarantees of the model
- Pros/Cons/Connections to other models

Predicting probabilities

According to the logistic regression model, we have

$$P(y = 1|\mathbf{x}; \mathbf{w}) = \sigma(\mathbf{w}^T\mathbf{x}) = \frac{1}{1 + \exp(-\mathbf{w}^T\mathbf{x})}$$

$$P(y = -1|\mathbf{x}; \mathbf{w}) = 1 - \sigma(\mathbf{w}^T\mathbf{x}) = \frac{\exp(-\mathbf{w}^T\mathbf{x})}{1 + \exp(-\mathbf{w}^T\mathbf{x})}$$

Predicting probabilities

According to the logistic regression model, we have

$$P(y = 1|\mathbf{x}; \mathbf{w}) = \sigma(\mathbf{w}^T\mathbf{x}) = \frac{1}{1 + \exp(-\mathbf{w}^T\mathbf{x})}$$

$$P(y = -1|\mathbf{x}; \mathbf{w}) = 1 - \sigma(\mathbf{w}^T\mathbf{x}) = \frac{\exp(-\mathbf{w}^T\mathbf{x})}{1 + \exp(-\mathbf{w}^T\mathbf{x})}$$

$$\frac{1}{1 + \exp(\mathbf{w}^T\mathbf{x})}$$

Predicting probabilities

According to the logistic regression model, we have

$$P(y = 1|\mathbf{x}; \mathbf{w}) = \sigma(\mathbf{w}^T\mathbf{x}) = \frac{1}{1 + \exp(-\mathbf{w}^T\mathbf{x})}$$

$$P(y = -1|\mathbf{x}; \mathbf{w}) = 1 - \sigma(\mathbf{w}^T\mathbf{x}) = \frac{1}{1 + \exp(\mathbf{w}^T\mathbf{x})}$$

Predicting probabilities

According to the logistic regression model, we have

$$P(y = 1|\mathbf{x}; \mathbf{w}) = \sigma(\mathbf{w}^T\mathbf{x}) = \frac{1}{1 + \exp(-\mathbf{w}^T\mathbf{x})}$$

$$P(y = -1|\mathbf{x}; \mathbf{w}) = 1 - \sigma(\mathbf{w}^T\mathbf{x}) = \frac{1}{1 + \exp(\mathbf{w}^T\mathbf{x})}$$

Or equivalently

$$P(y|\mathbf{x}; \mathbf{w}) = \frac{1}{1 + \exp(-y\mathbf{w}^T\mathbf{x})}$$

Predicting a label

$$P(y = 1|\mathbf{x}; \mathbf{w}) = \sigma(\mathbf{w}^T\mathbf{x}) = \frac{1}{1 + \exp(-\mathbf{w}^T\mathbf{x})}$$

- Compute P(y =1 | x; w)

- If this is greater than half, predict 1 else predict -1
  - What does this correspond to in terms of $\mathbf{w}^T\mathbf{x}$?

Decision Boundary

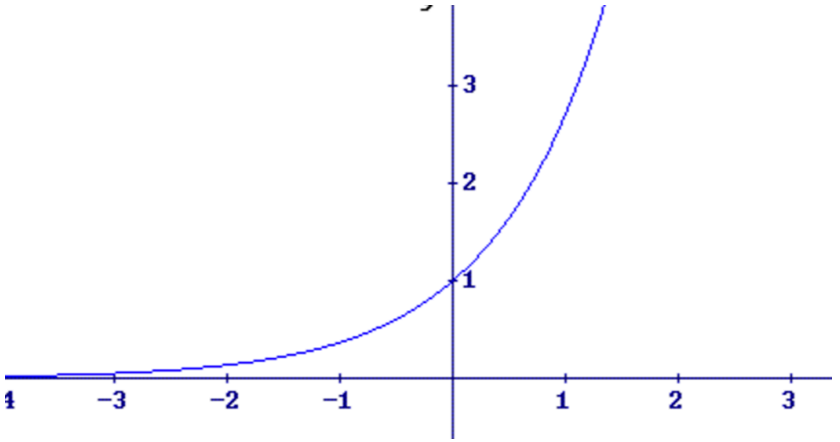
- Predict y=1 if P(y=1|x,w) ≥ P(y= -1|x, w)

- When does this happen?
  - $\frac{1}{1+\exp(-w^T x)} \geq 0.5$

    $\Rightarrow 1 + \exp(-w^T x) \leq 2$

    $\Rightarrow \exp(-w^T x) \leq 1$

    $\Rightarrow w^T x \geq 0$



DecisionType equation here. boundary

$$P(y=1|\mathbf{x},\mathbf{w}) = \frac{e^{\mathbf{w}^T\mathbf{x}}}{1+e^{\mathbf{w}^T\mathbf{x}}} = \frac{1}{1+e^{-\mathbf{w}^T\mathbf{x}}}$$

$$P(y=-1|\mathbf{x},\mathbf{w}) = \frac{1}{1+e^{\mathbf{w}^T\mathbf{x}}}$$

$$\log \frac{P(Y=1 \mid x)}{P(Y=-1 \mid x)} = w^T x$$

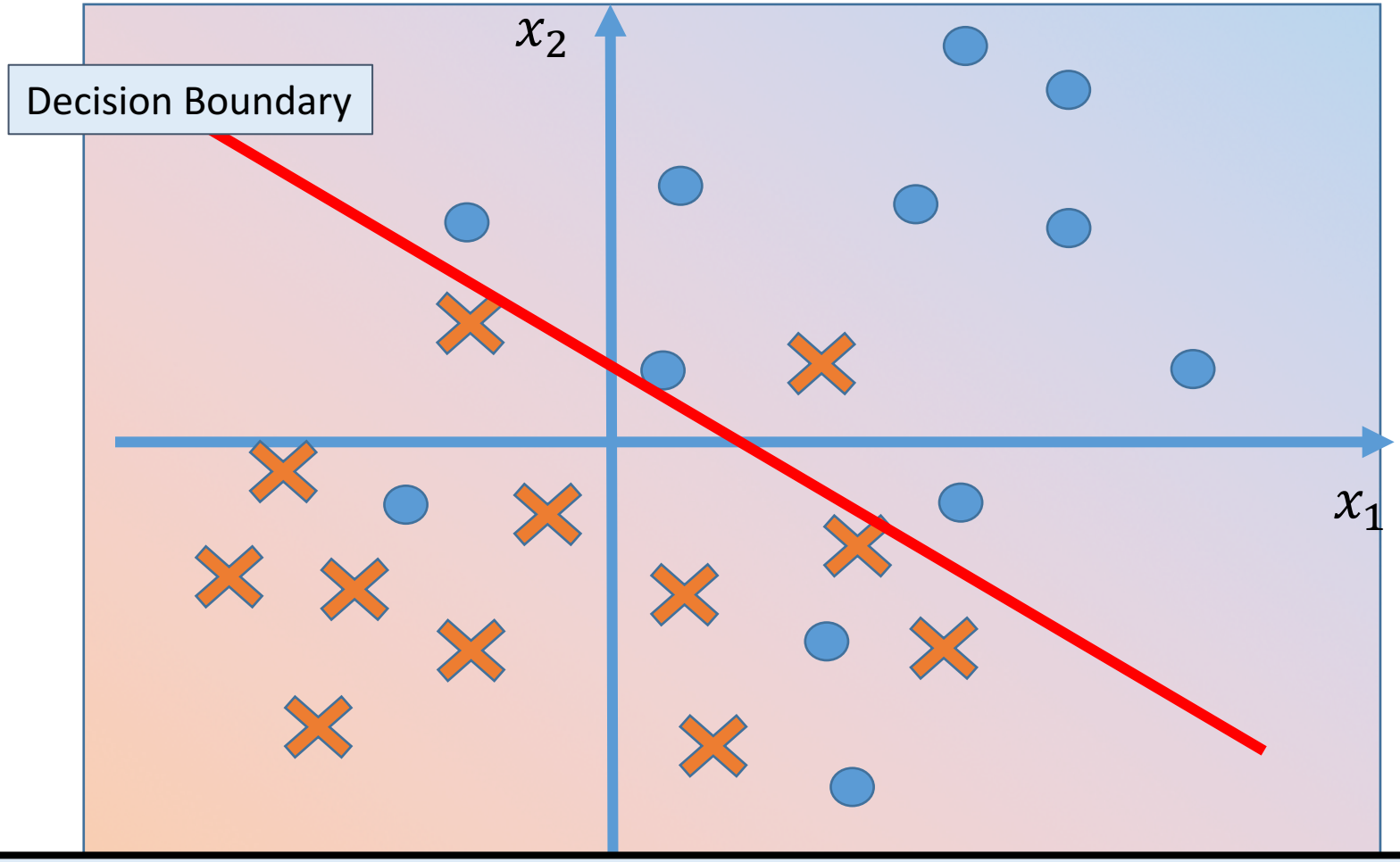
$\log(z) \geq 0 \quad iff \quad z \geq 1$

- The decision boundary?

$$w^T x = 0$$

- $\sigma(w^T x)$ is non-linear but the decision boundary is linear!

Prediction by logistic regression



A probabilistic model of modeling $\sigma(w^T x + b) = P(y = 1|x)$

How to train a logistic regression model?

Logistic Regression: Setup

- The setting
  - Binary classification
  - Inputs: Feature vectors $x \in R^N$
  - Labels: $\mathbf{y} \in \{-1, +1\}$
- Training data
  - $S = \{(\mathbf{x}_i, \mathbf{y}_i)\}$, m examples
- Hypothesis space

$$H = \{\ \mathrm{h} \mid h : X \rightarrow P(Y \mid X), h(x) = \sigma\ (\ w^T x + b)\ \}$$

$$\sigma(z) = \frac{1}{1 + \exp(-z)}$$

Logistic Regression: Setup

- Training data
  - S = $\{(\mathbf{x}_i, \mathbf{y}_i)\}$, m examples
- Hypothesis space

$$H = \{\ \mathrm{h} \mid h : X \to Y, h(x) = \sigma\ (\ w^T x + b)\ \}$$

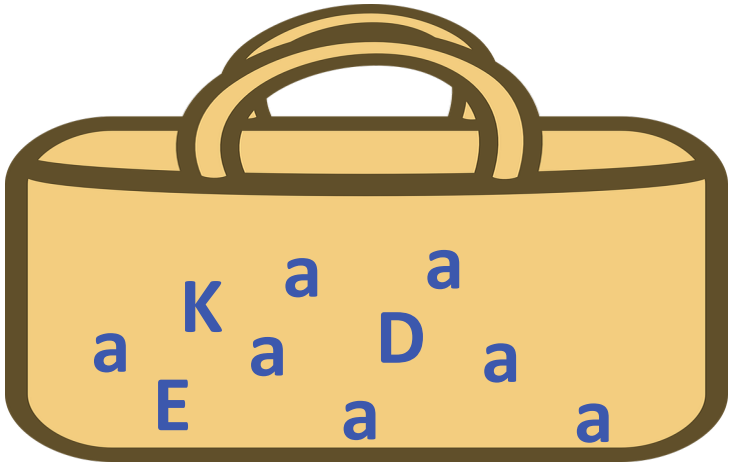
$$\sigma(z) = \frac{1}{1 + \exp(-z)}$$

- Learning Goal:
  - Find an $\mathrm{h} \in H$, such that $\mathrm{h}(\mathrm{x}) \approx P(y = 1 \mid x)$
  - How to?

Maximum Likelihood

- Which bag of words more likely generate:

  aDaaa

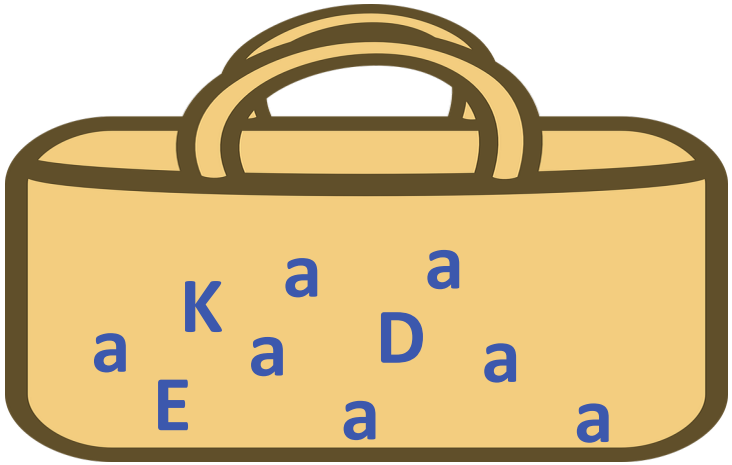


Maximum Likelihood

- Which bag of words more likely generate:

  aDaaa

$0.7 \times 0.1 \times 0.7 \times 0.7 \times 0.7$
$= 2.401 \times 10^{-2}$



$0.2 \times 0.1 \times 0.2 \times 0.2 \times 0.2$
$= 1.6 \times 10^{-4}$

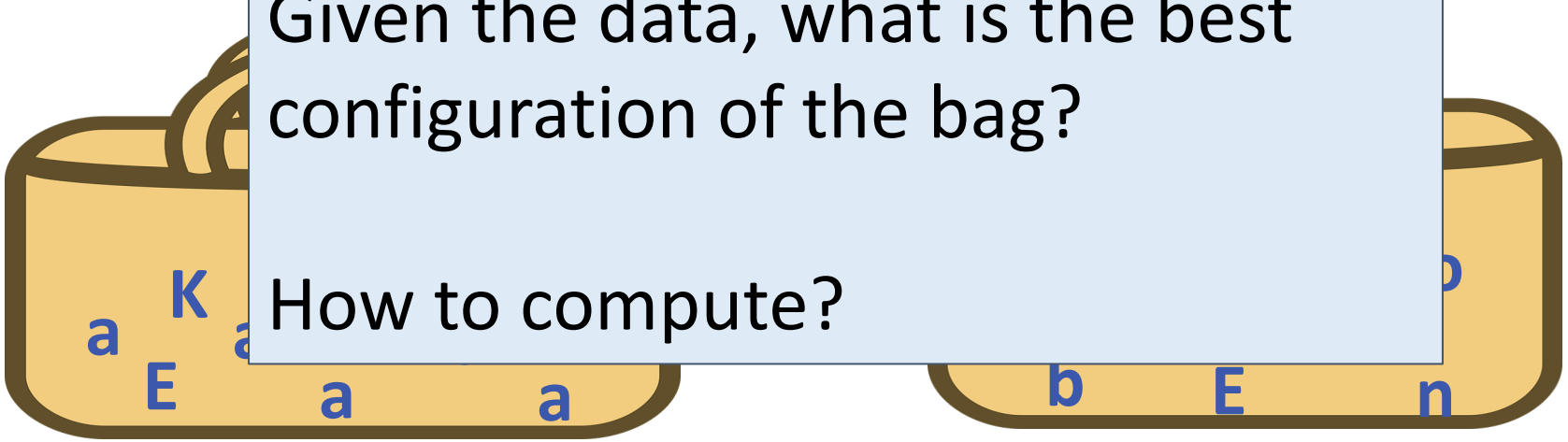
Maximum Likelihood

- Which bag of words more likely generate:

  aDaaa

$0.7 \times 0.1 \times 0.7 \times 0.7 \times 0.7 \quad 0.3 \times 0.1 \times 0.3 \times 0.2 \times 0.2$

$= 2.401$



Given the data, what is the best configuration of the bag?

How to compute?

Drawing color cards from the envelope

- Let say we have several cards in the envelope
- Assume

$P(card = yellow) = \theta$

$P(card = purple) = 1 - \theta$

Drawing color cards from the envelope

- Sample with replacement n times
- k times we get yellow card, and n-k times, we get purple card
- The joint probability (likelihood)

$$\theta^k(1-\theta)^{n-k}$$

- What is the best p making the joint probability maximal?

Drawing color cards from the envelope

- Solving $\max_\theta \theta^k(1-\theta)^{n-k}$
- Equivalently, we can solve

$$\max_\theta \log(\theta^k(1-\theta)^{n-k})$$

$$\max_\theta \; \mathrm{k} \; \log \; \theta + (n-k)\log(1-\theta)$$

- At the optimum,

$$\frac{\mathrm{d}(\mathrm{k} \; \log \; \theta + (n-k)\log(1-\theta))}{d\theta} = 0$$

The usual trick: Convert products to sums by taking log

Recall that this works only because log is an increasing function and the maximizer will not change

Drawing color cards from the envelope

$$\frac{\mathrm{d}(\mathrm{k}\ \log\ \theta+(n-k)\log(1-\theta))}{d\theta}=0$$

$$\Rightarrow \frac{\mathrm{k}}{\theta}-\frac{n-\mathrm{k}}{1-\theta}=0$$

For this simple problem, we have a closed-form solution.
We are not always lucky like this

$$\Rightarrow\ \theta(\mathrm{n}-\mathrm{k})=(1-\theta)\mathrm{k}$$

$$\Rightarrow \theta=\frac{\mathrm{k}}{\mathrm{n}}$$

Maximum Likelihood Estimator (formal definition)

- Likelihood function of parameters

Let $X_1, \ldots, X_N$ be IID (independent and identically distributed) with PDF $p(x|\theta)$ (also written as $p(x;\theta)$). The *likelihood function* is defined by $L(\theta)$,

$$L(\theta) = p(X_1, \ldots, X_N; \theta). \qquad = \prod_{i=1}^{N} p(X_i; \theta).$$

Notes The likelihood function is just the joint density of the data, except that we treat it as a function of the parameter θ.

Maximum Likelihood Estimation

❖ Maximum Likelihood Estimator

Definition: The maximum likelihood estimator (MLE) $\hat{\theta}$, is the value of θ that maximizes $L(\theta)$.

The log-likelihood function is defined by $l(\theta) = \log L(\theta)$. Its maximum occurs at the same place as that of the likelihood function.

Tricks for computation: Log-Likelihood

The log-likelihood function is defined by $l(\theta) = \log L(\theta)$. Its maximum occurs at the same place as that of the likelihood function.

- Using logs simplifies mathemetical expressions (converts exponents to products and products to sums)
- Using logs helps with numerical stabilitity

The same is true of the likelihood function times any constant. Thus we shall often drop constants in the likelihood function.

Back to Logistic regression

- Training data
  - S = $\{(\mathbf{x}_i, \mathbf{y}_i)\}$, m examples
- Hypothesis space

$$H = \{\ \mathrm{h} \mid h : X \to Y, h(x) = \sigma\ (\ w^T x + b)\ \}$$

$$\sigma(z) = \frac{1}{1 + \exp(-z)}$$

- Learning Goal:
  - Find an $\mathrm{h} \in H$, such that $\mathrm{h}(\mathrm{x}) \approx P(y = 1 \mid x)$
  - How to? Maximum Likelihood Estimator

Maximum likelihood estimator for logistic regression

$$\underset{\mathbf{w}}{\operatorname{argmax}} P(S|\mathbf{w}) = \underset{\mathbf{w}}{\operatorname{argmax}} \prod_{i=1}^{m} P(y_i|\mathbf{x}_i, \mathbf{w})$$

Maximum likelihood estimator for logistic regression

$$\underset{\mathbf{w}}{\operatorname{argmax}} P(S|\mathbf{w}) = \underset{\mathbf{w}}{\operatorname{argmax}} \prod_{i=1}^{m} P(y_i|\mathbf{x}_i, \mathbf{w})$$

Equivalent to solving

$$\max_{\mathbf{w}} \sum_{i}^{m} \log P(y_i|\mathbf{x}_i, \mathbf{w})$$

Maximum likelihood estimator for logistic regression

$$\underset{\mathbf{w}}{\operatorname{argmax}} P(S|\mathbf{w}) = \underset{\mathbf{w}}{\operatorname{argmax}} \prod_{i=1}^{m} P(y_i|\mathbf{x}_i, \mathbf{w})$$

$$\max_{\mathbf{w}} \sum_{i}^{m} \log P(y_i|\mathbf{x}_i, \mathbf{w})$$

By definition, we know that

$$P(y|\mathbf{w}, \mathbf{x}) = \sigma(y_i \mathbf{w}^T \mathbf{x}_i) = \frac{1}{1 + \exp(-y_i \mathbf{w}^T \mathbf{x}_i)}$$

Maximum likelihood estimator for logistic regression

$$\underset{\mathbf{w}}{\operatorname{argmax}} P(S|\mathbf{w}) = \underset{\mathbf{w}}{\operatorname{argmax}} \prod_{i=1}^{m} P(y_i|\mathbf{x}_i, \mathbf{w})$$

$$\max_{\mathbf{w}} \sum_{i}^{m} \log P(y_i|\mathbf{x}_i, \mathbf{w})$$

Equivalent to solving

$$\max_{\mathbf{w}} \sum_{i}^{m} -\log(1 + \exp(-y_i \mathbf{w}^T \mathbf{x}_i))$$

$$P(y|\mathbf{w}, \mathbf{x}) = \frac{1}{1 + \exp(-y\, \mathbf{w}^T \boldsymbol{x})}$$

Maximum likelihood estimator for logistic regression

$$\underset{\mathbf{w}}{\operatorname{argmax}} P(S|\mathbf{w}) = \underset{\mathbf{w}}{\operatorname{argmax}} \prod_{i=1}^{m} P(y_i|\mathbf{x}_i, \mathbf{w})$$

$$\max_{\mathbf{w}} \sum_{i}^{m} \log P(y_i|\mathbf{x}_i, \mathbf{w})$$

Equivalent to solving

$$\max_{\mathbf{w}} \sum_{i}^{m} -\log(1 + \exp(-y_i \mathbf{w}^T \mathbf{x}_i))$$

This is somehow ugly

Minimizing Negative Log-likelihood

- It is convenient to work with its negation termed negative log likelihood

$$\max_{\mathbf{w}} \sum_{i}^{m} -\log(1 + \exp(-y_i \mathbf{w}^T \mathbf{x}_i))$$

$$\min_{\mathbf{w}} \sum_{i}^{m} \log(1 + \exp(-y_i \mathbf{w}^T \mathbf{x}_i))$$

Alternative Formulation

- Sometimes, we define
  - Inputs: Feature vectors $x \in R^N$
  - Labels: $y \in \{0, 1\}$ (or $y \in [0, 1]$)
- Training data
  - S = $\{(\mathbf{x}_i, \mathbf{y}_i)\}$, m examples
- Hypothesis space

$$H = \{ \text{h} \mid h : X \to Y, h(x) = \sigma (w^T x + b) \}$$

$$\sigma(z) = \frac{1}{1 + \exp(-z)}$$

Alternative representation

$$P(y|\mathbf{w},\mathbf{x}) = \boldsymbol{\sigma}\left(\boldsymbol{w^T x}\right) = \frac{1}{1+\exp(-\mathrm{y}\ \ \mathbf{w}^T\mathbf{x}_i)}$$

Maximum Likelihood estimator:

$$\underset{\mathbf{w}}{\operatorname{argmax}}\, P(S|\mathbf{w}) = \underset{\mathbf{w}}{\operatorname{argmax}} \prod_{i=1}^{m} P(y_i|\mathbf{x}_i,\mathbf{w})$$

$$P(y_i|\mathbf{x}_i,\mathbf{w}) = \begin{cases} \sigma(w^T x) & if\ y_i = 1 \\ 1-\sigma(w^T x) & if\ y_i = 0 \end{cases}$$

Alternative representation

$$P(y|\mathbf{w},\mathbf{x}) = \boldsymbol{\sigma}\left(\boldsymbol{w^T x}\right) = \frac{1}{1+\exp(-\mathrm{y}\,\mathbf{w}^T\mathbf{x}_i)}$$

Maximum Likelihood estimator:

$$\underset{\mathbf{w}}{\operatorname{argmax}}\, P(S|\mathbf{w}) = \underset{\mathbf{w}}{\operatorname{argmax}} \prod_{i=1}^{m} P(y_i|\mathbf{x}_i,\mathbf{w})$$

Equivalent to solving

$$P(y|\mathbf{x}_i,\mathbf{w}) = \begin{cases} \sigma(w^T x) & if\ y=1 \\ 1-\sigma(w^T x) & if\ y=0 \end{cases}$$

$$\underset{\mathbf{w}}{\operatorname{argmax}} \prod_{i=1}^{m} P(y=1|\mathbf{x}_i,\mathbf{w})^{y_i} P(y=0|\mathbf{x}_i,\mathbf{w})^{1-y_i}$$

Alternative representation

$$P(y|\mathbf{w},\mathbf{x}) = \boldsymbol{\sigma}\left(\boldsymbol{w^T x}\right) = \frac{1}{1+\exp(-\mathrm{y}\ \mathbf{w}^T\mathbf{x}_i)}$$

Maximum Log-Likelihood estimator:

$$\underset{\mathbf{w}}{\operatorname{argmax}} \prod_{i=1}^{m} P(y=1|\mathbf{x}_i,\mathbf{w})^{y_i} P(y=0|\mathbf{x}_i,\mathbf{w})^{1-y_i}$$

Equivalent to solving

$$P(y|\mathbf{x}_i,\mathbf{w}) = \begin{cases} \sigma(w^T x) & if\ y=1 \\ 1-\sigma(w^T x) & if\ y=0 \end{cases}$$

$$\underset{\mathbf{w}}{\operatorname{argmax}} \sum_{i=1}^{m} y_i \log\sigma(w^T x) + (1-y_i)\log(1-\sigma(w^T x))$$

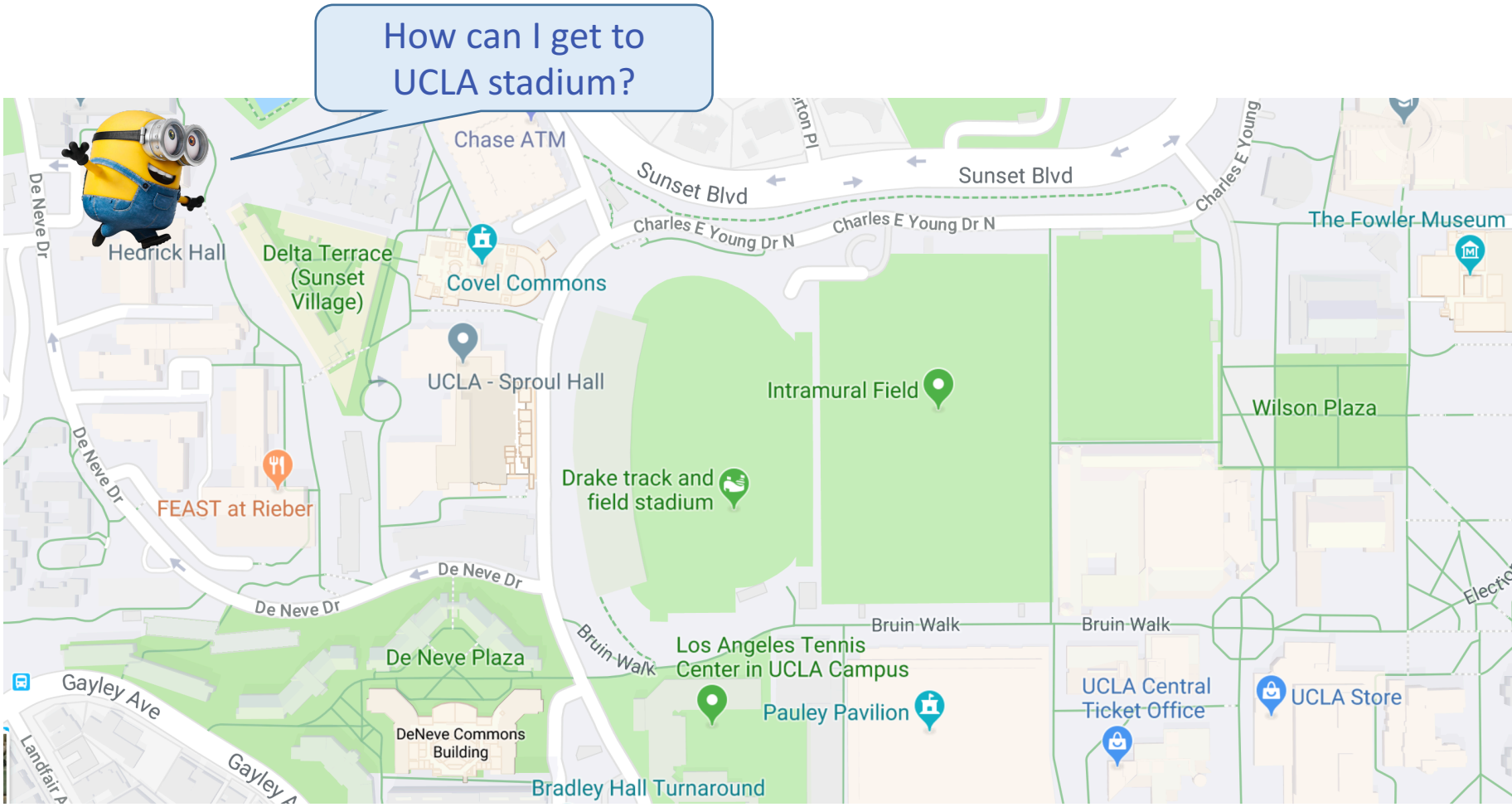
How to solve this?

$$\min_{\mathbf{w}} \sum_{i}^{m} \log(1 + \exp(-y_i \mathbf{w}^T \mathbf{x}_i))$$

- There is no closed-form solution
- One way to solve it is by gradient descent

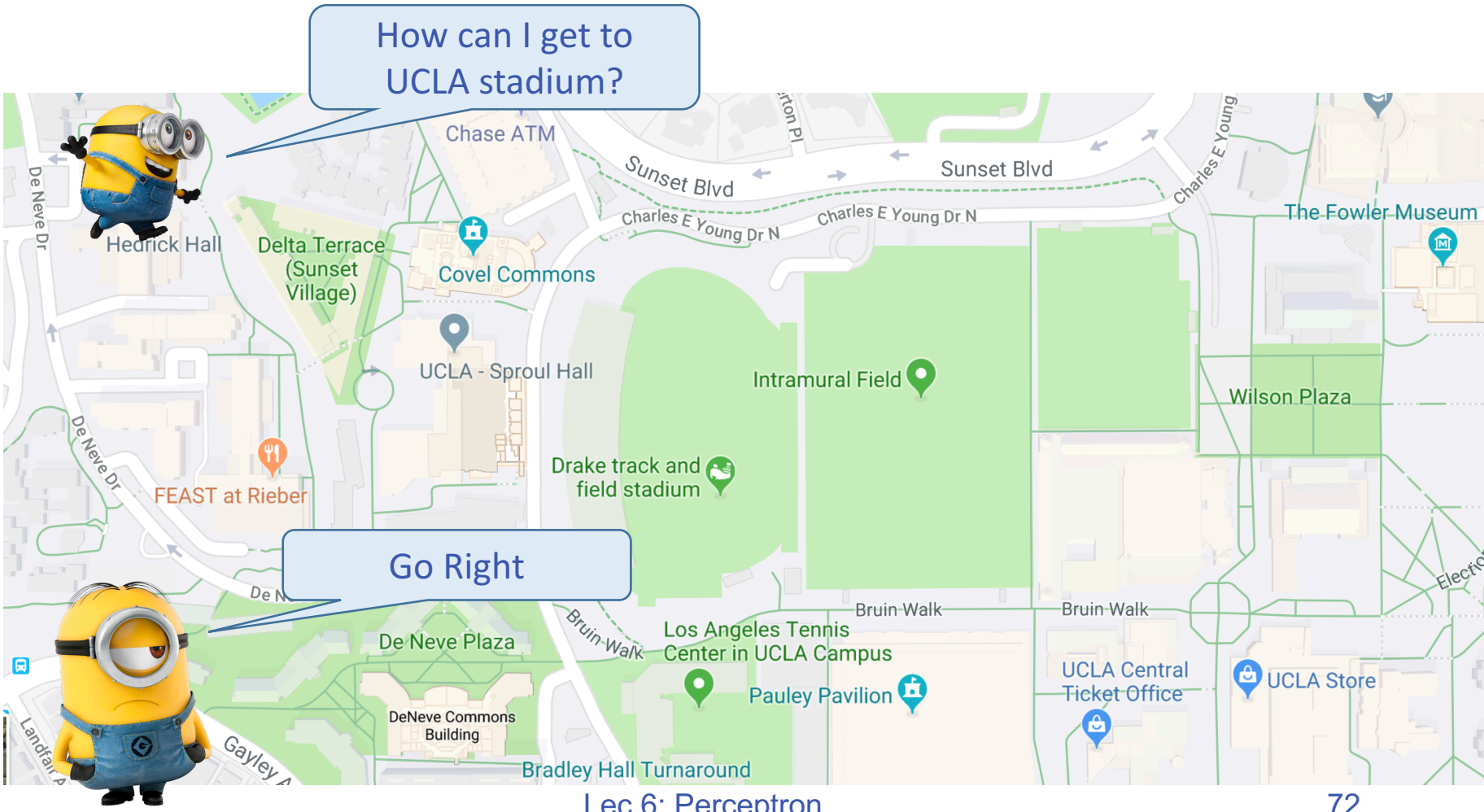
Intuition

- Asking direction



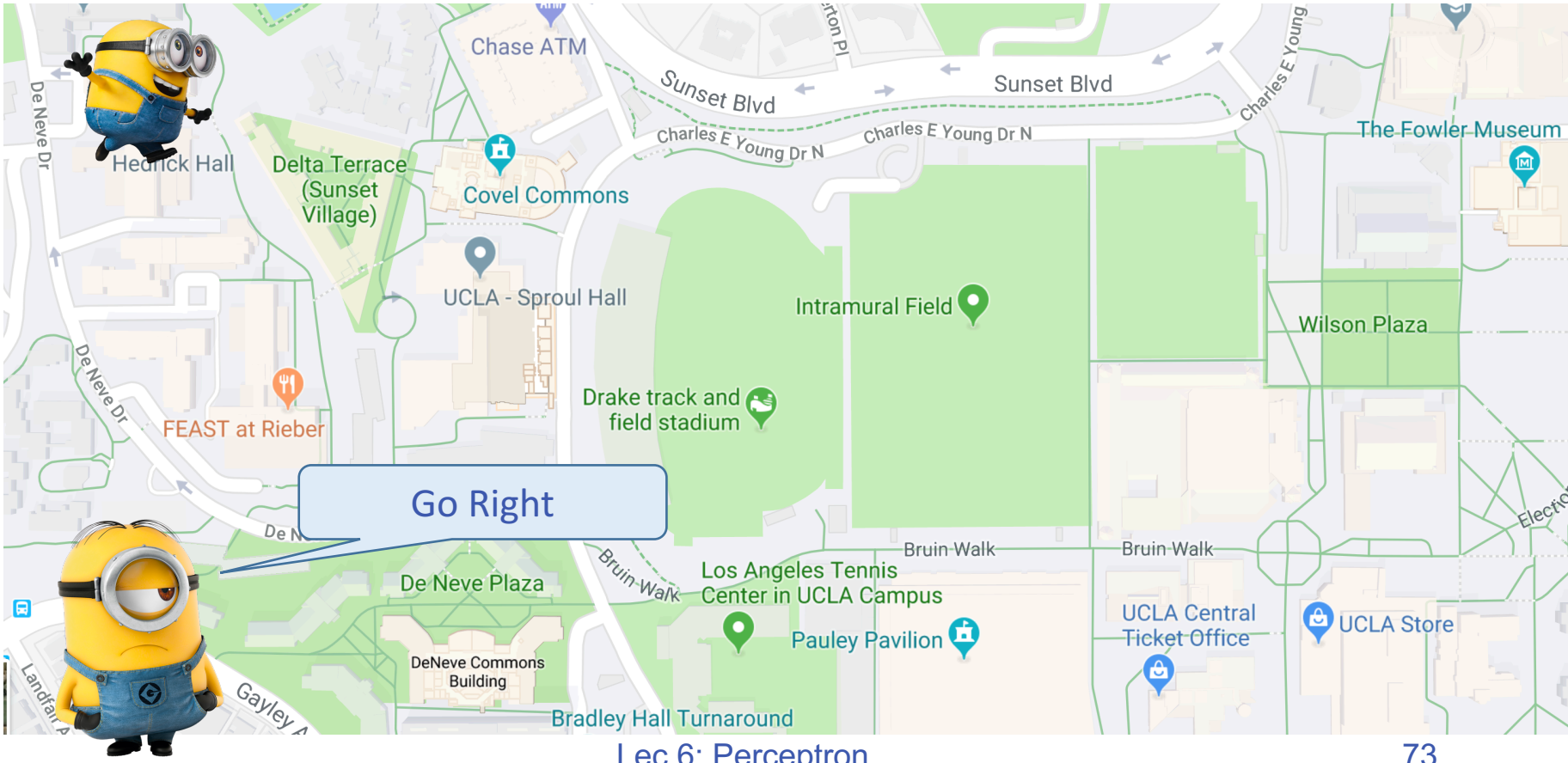
Intuition

- Asking direction



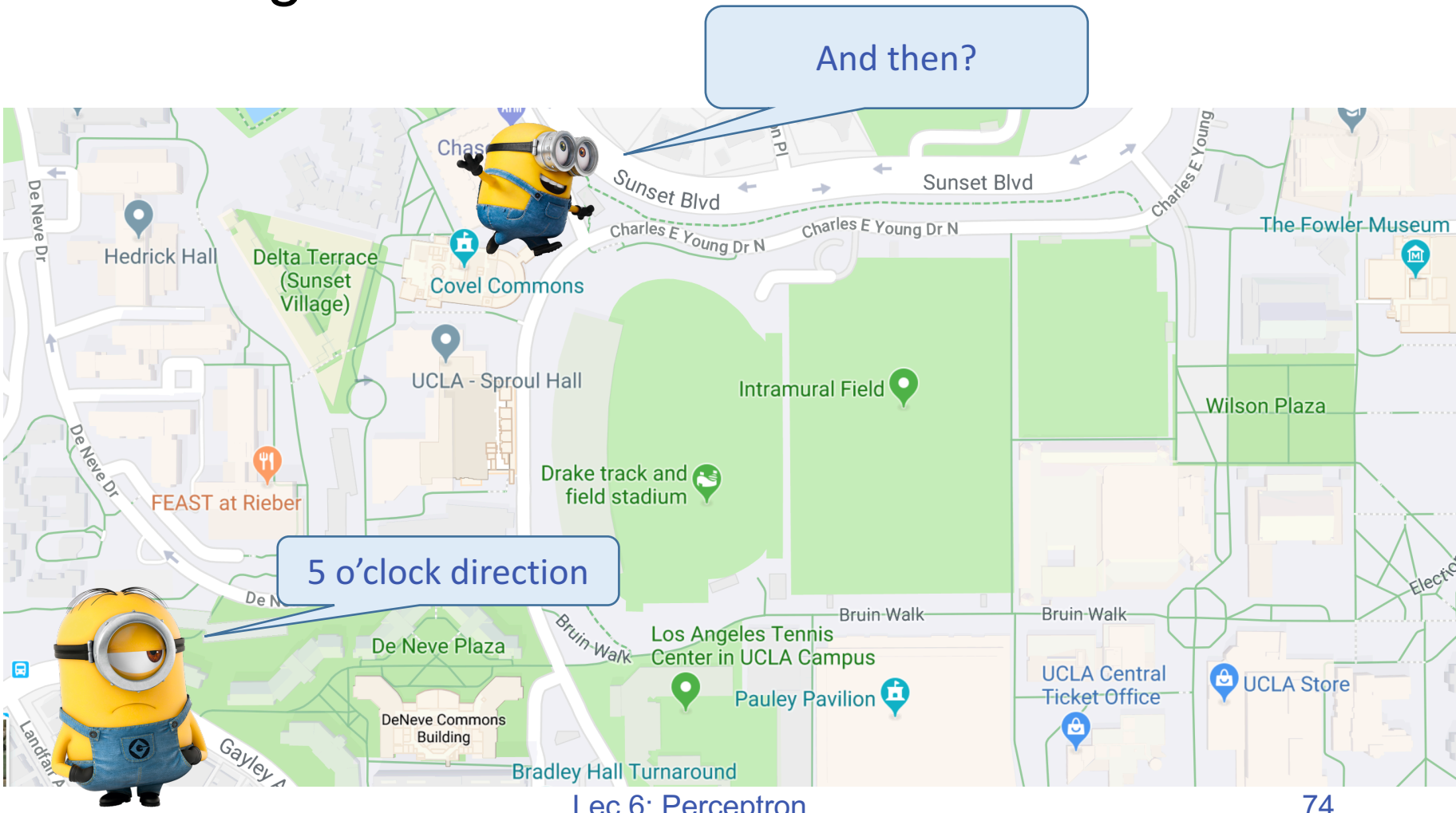
Intuition

- Asking direction



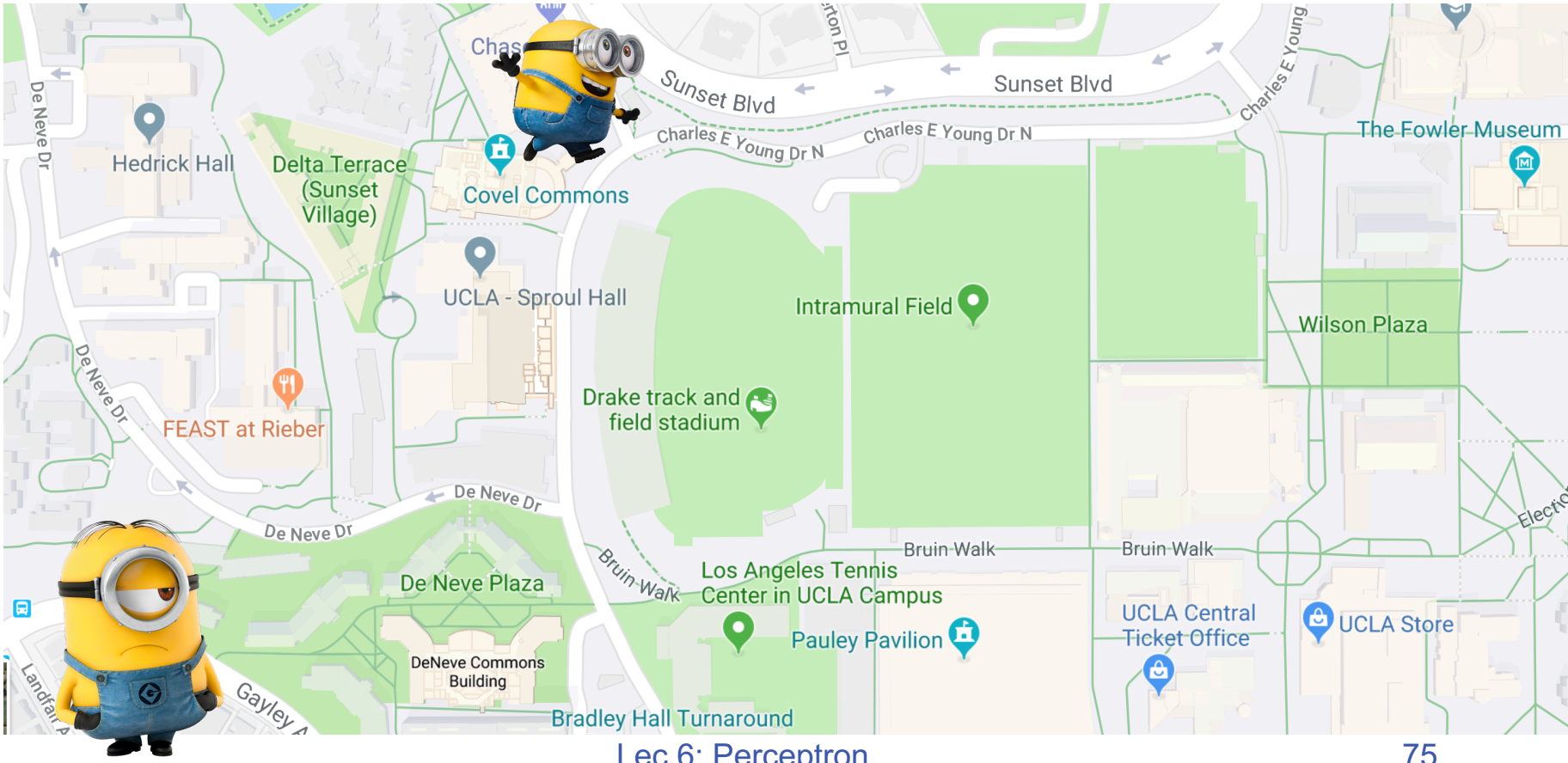
Intuition

- Asking direction



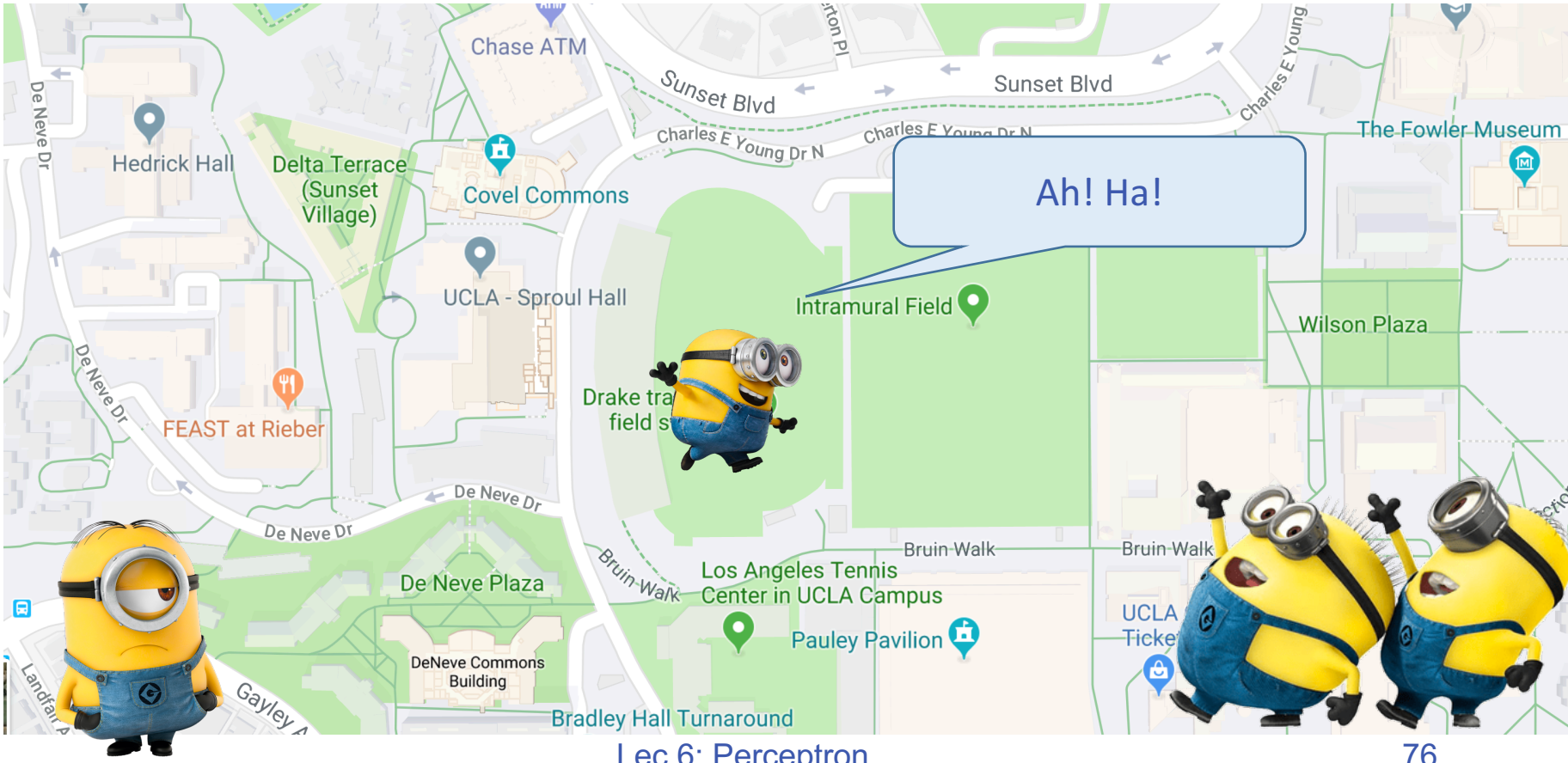
Intuition

- Asking direction



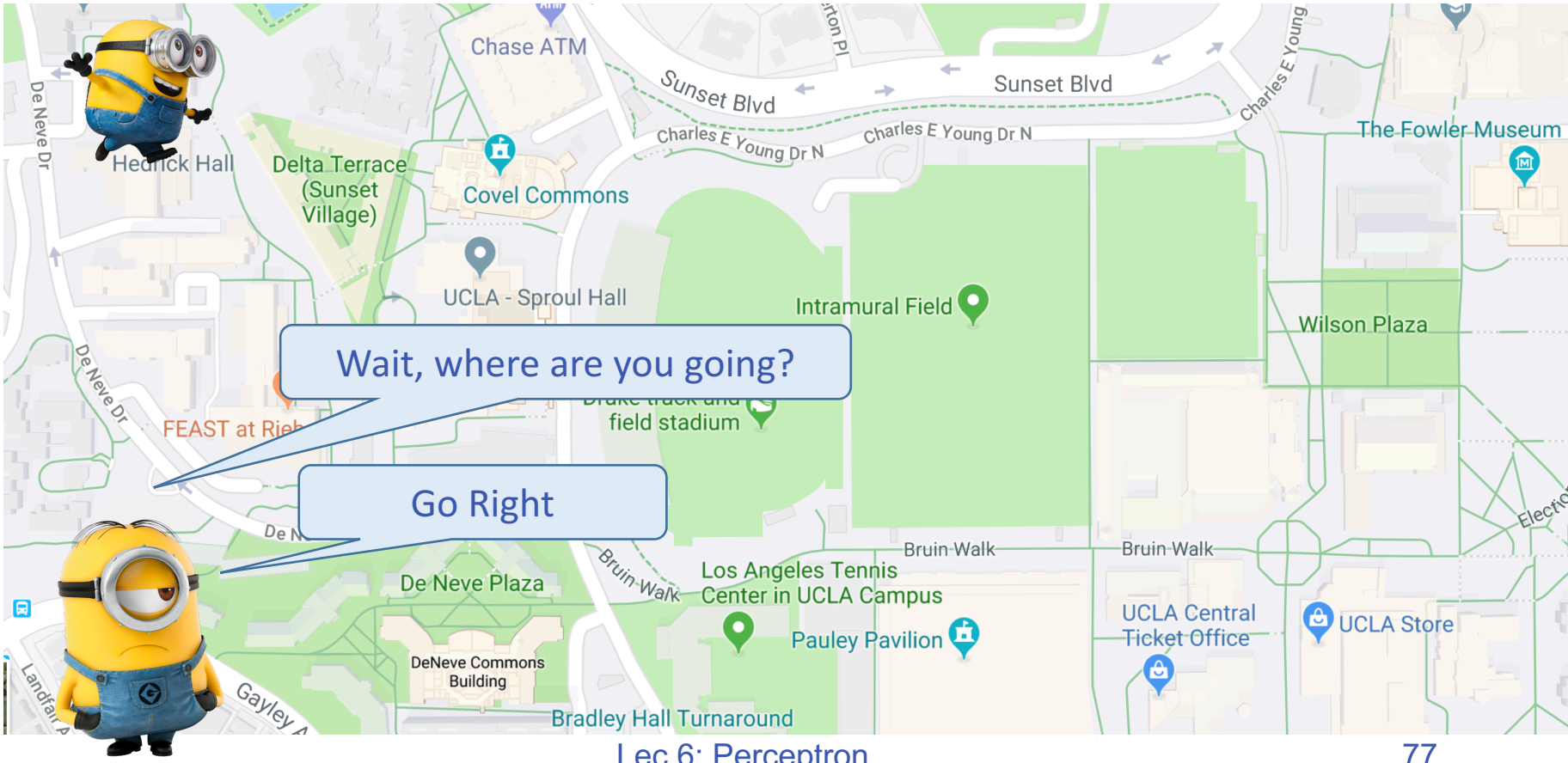
Intuition

- Asking direction



Intuition

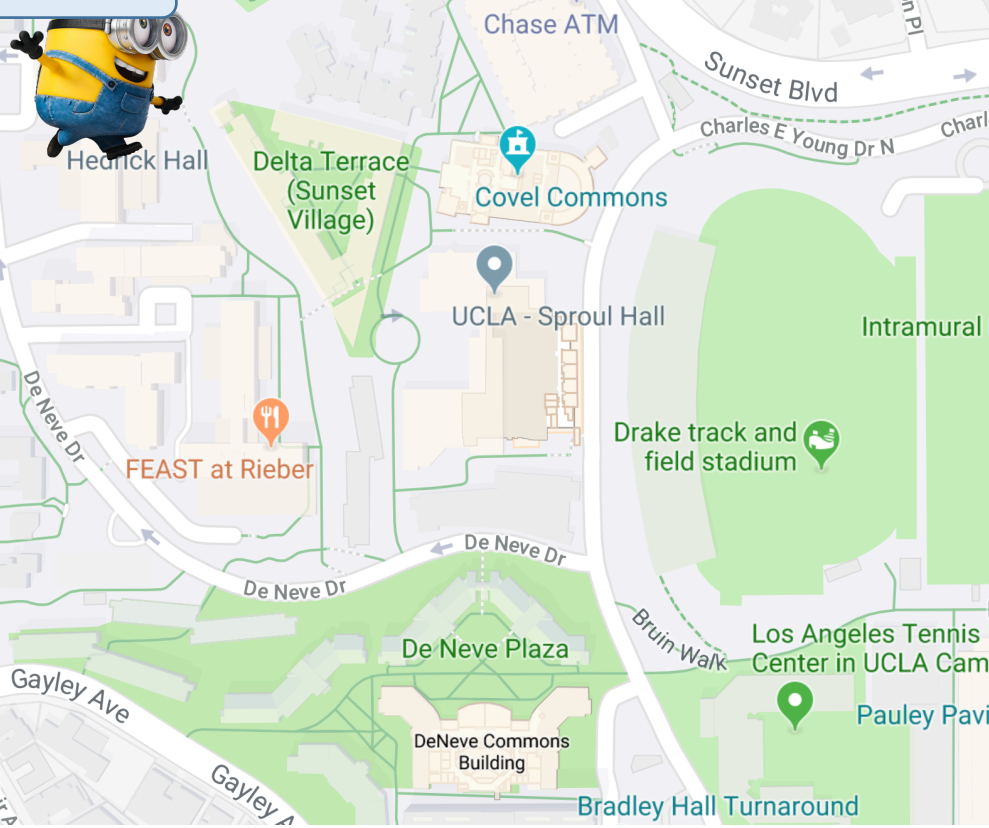
- What may go wrong? Incorrect Step-size



Intuition

- What may go wrong? Incorrect Direction

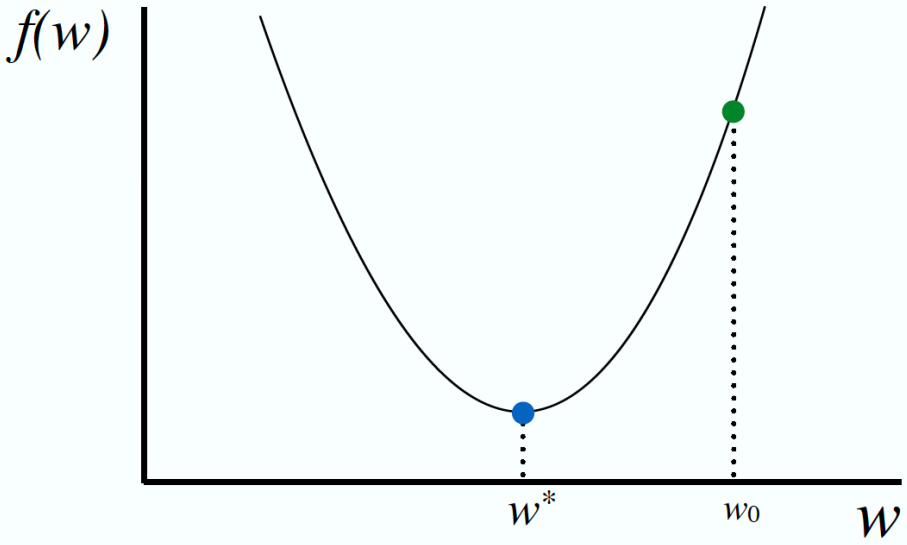
AHHhhhhhhhh….



Go left, hehehe…

Gradient descent

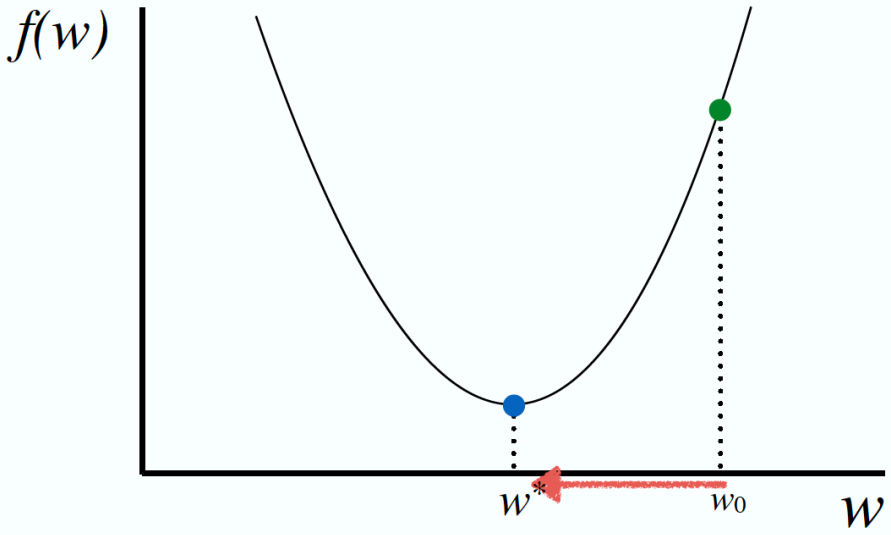
Start at a random point



Gradient descent

Start at a random point

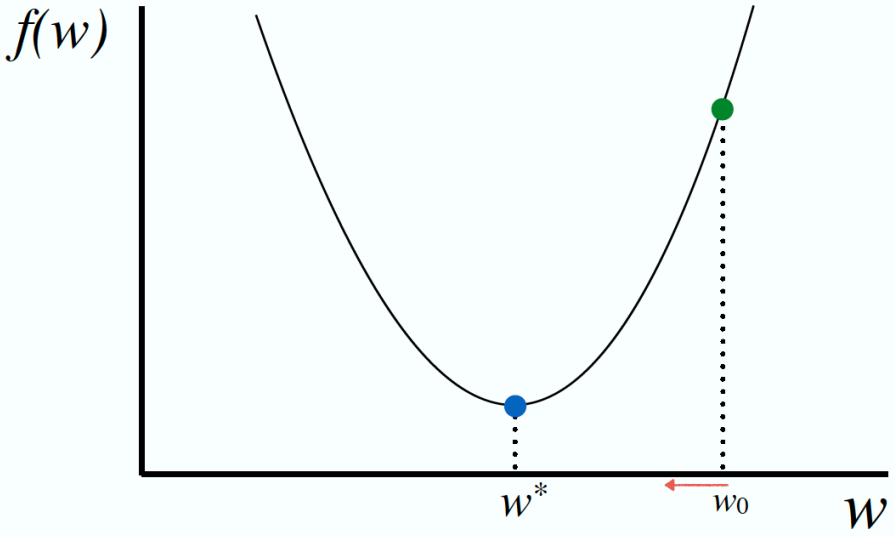
Determine a descent direction



Gradient descent

Start at a random point

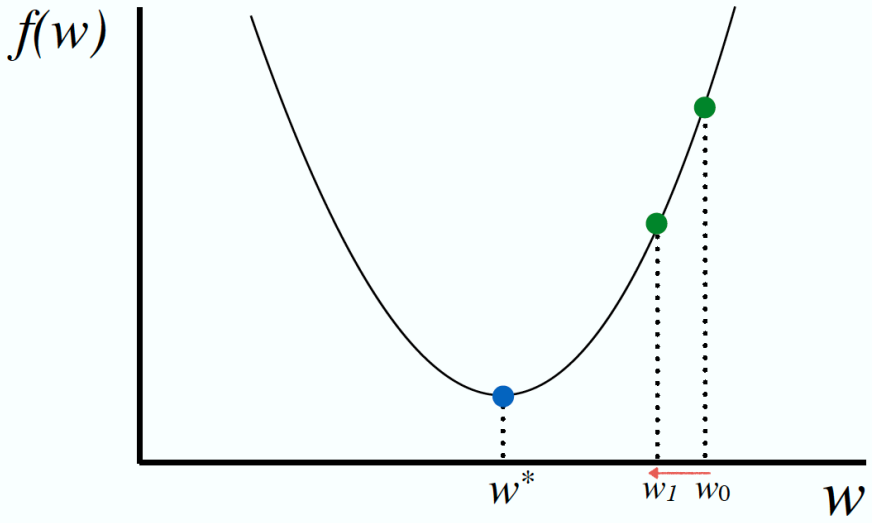
Determine a descent direction
Choose a step size



Gradient descent

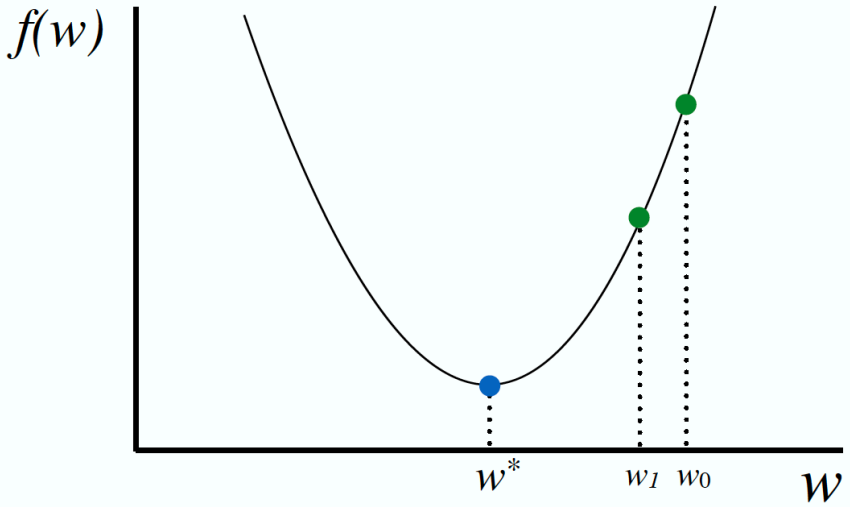
Start at a random point

Determine a descent direction
Choose a step size
Update



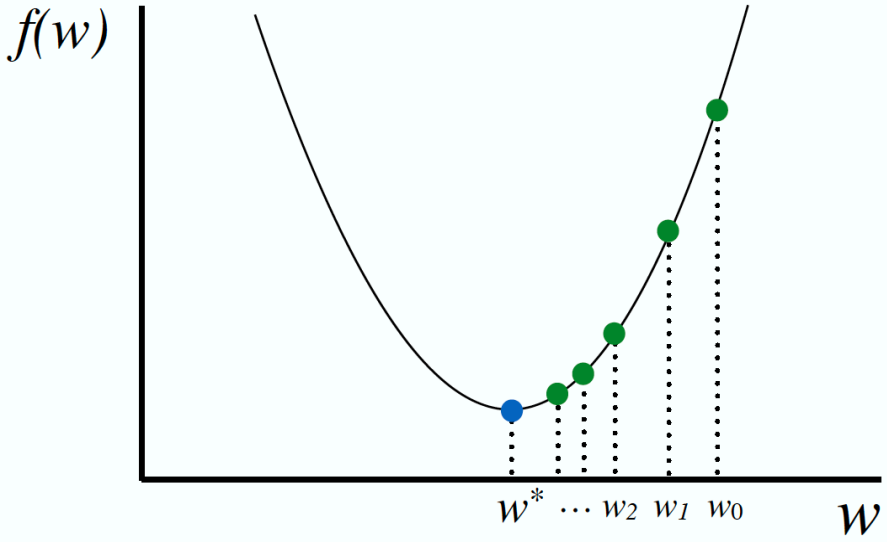
Gradient descent

Start at a random point
Repeat
    Determine a descent direction
    Choose a step size
    Update
Until stopping criterion is satisfied



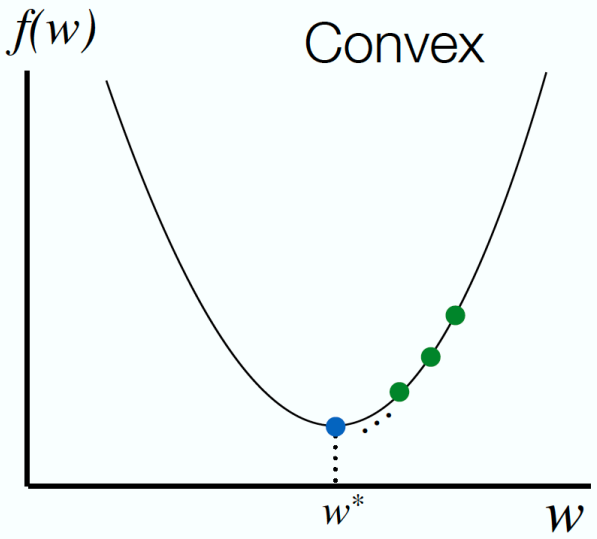
Gradient descent

Start at a random point
Repeat
    Determine a descent direction
    Choose a step size
    Update
Until stopping criterion is satisfied

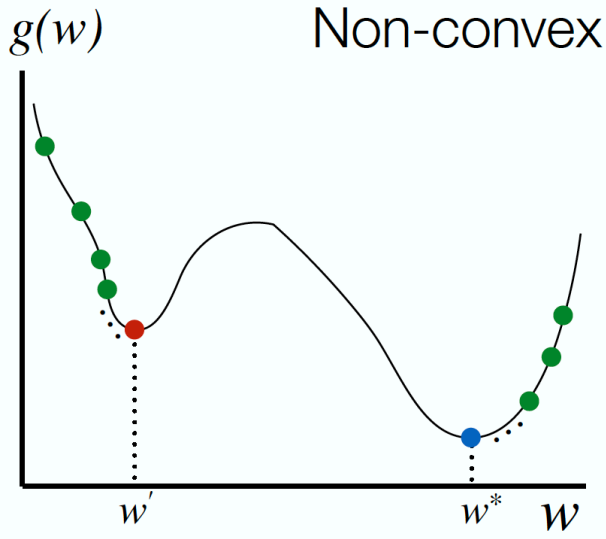


Where will we converge?

- If function is convex, it converges to the global optimum (need proper choice of step-size)



Any local minimum is a global minimum



Multiple local minima may exist

Exercise

- Let $\sigma(z) = \frac{1}{1+\exp(-z)}$, show

  $\sigma(-z) = 1 - \sigma(z)$

- What is the gradient of

$$\underset{\mathbf{w}}{\operatorname{argmax}} \sum_{i=1}^{m} y_i \log \sigma(w^T x) + (1 - y_i) \log(1 - \sigma(w^T x))$$