CS 540 Introduction to Artificial Intelligence
Linear Models & Linear Regression

Yudong Chen
University of Wisconsin-Madison

Oct 12, 2021

Announcements

- **Homeworks:**
  - HW5: due next Tuesday

- Class roadmap:

Thursday, Oct 7	ML Unsupervised II
Tuesday, Oct 12	**ML Linear Regression**
Thursday, Oct 14	ML: Naïve Bayes, KNN
Tuesday, Oct 19	ML: Neural Networks I

Machine Learning

Outline

NOT TODAY

- Unsupervised Learning: Density Estimation
  - Kernel density estimation: high-level intro
- Supervised Learning & Linear Models
  - Parameterized model, model classes, linear models, train vs. test
- Linear Regression
  - Least squares, normal equations, residuals, logistic regression

Short Intro to Density Estimation

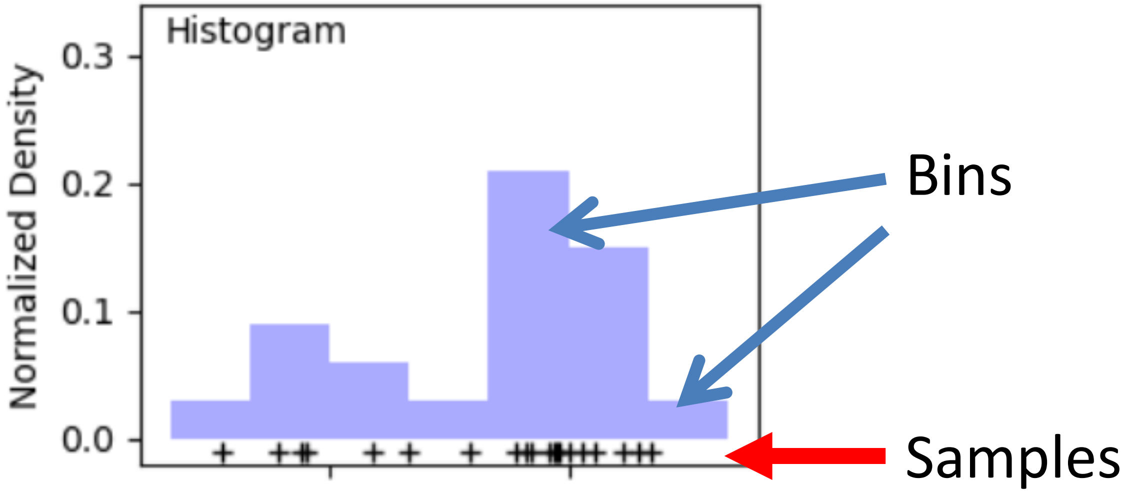
Goal: given samples $x_1, \ldots, x_n$ from some distribution *P*, estimate *P*.

- Compute statistics (mean, variance)
- Generate samples from *P*
- Run inference

Zach Monge

Simplest Idea: Histograms

Goal: given samples $x_1, \ldots, x_n$ from some distribution P, estimate P.



Define bins; count # of samples in each bin, normalize

Simplest Idea: Histograms

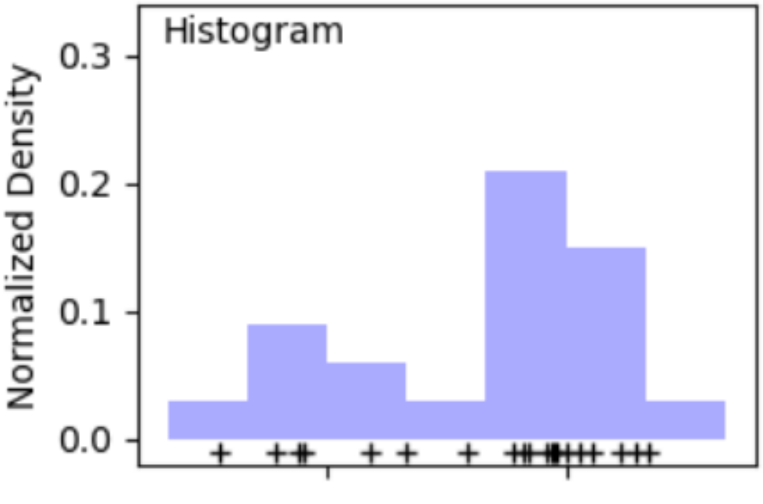
Goal: given samples x_1, …, x_n from some distribution *P*, estimate *P*.

Downsides:

i) High-dimensions: most bins empty

ii) Not continuous

iii) How to choose bins?



Kernel Density Estimation

Goal: given samples $x_1, \ldots, x_n$ from some distribution *P*, estimate *P*.

Idea: represent density as combination of "kernels"

$$f(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

Center at each point

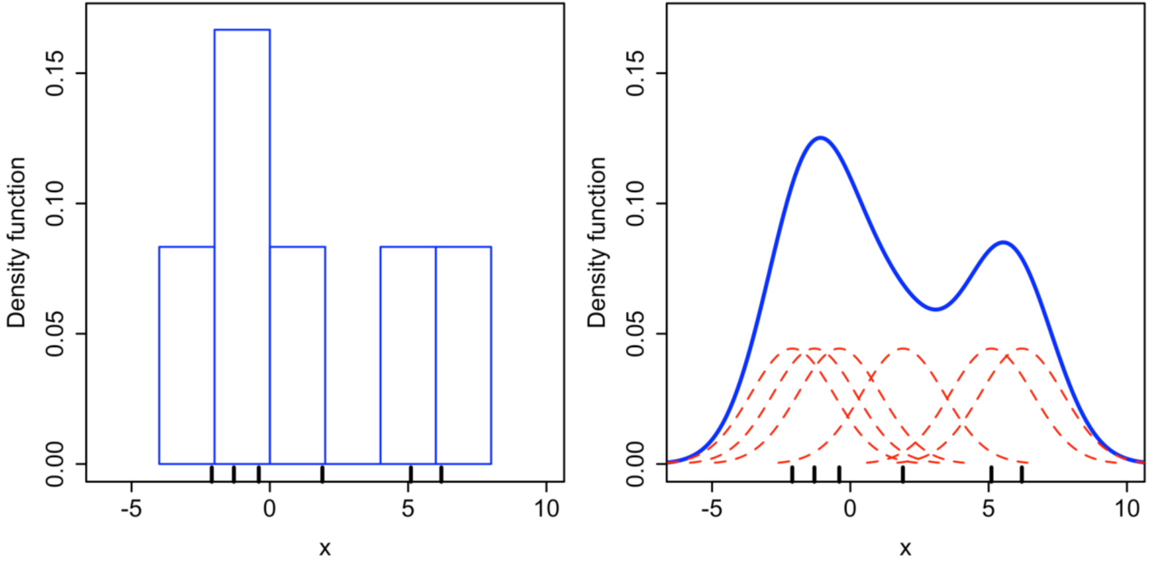
Kernel function: often Gaussian

Width parameter

Kernel Density Estimation

Idea: represent density as combination of kernels

- "Smooth" out the histogram



Break & Quiz

Q 1.1: Which of the following is **not** true?

- A. Using a Gaussian kernel for KDE, all possible values for x will have non-zero probability $f(x)$.
- B. The goal of KDE is to approximate the true probability distribution function of x.
- C. KDE cannot be applied if the data $x_1, \ldots, x_n$ are vectors
- D. With some kernels, KDE can assign zero probability to some subset of values for x.

Break & Quiz

Q 1.1: Which of the following is **not** true?

- A. Using a Gaussian kernel for KDE, all possible values for x will have non-zero probability $f(x)$. (Gaussian PDF positive for all inputs)
- B. The goal of KDE is to approximate the true probability distribution function of x. (same goal as histograms)
- **C. KDE cannot be applied if the data $x_1, \ldots, x_n$ are vectors**
- D. With some kernels, KDE can assign zero probability to some subset of values for x. (Consider K = uniform(0,1))

Back to Supervised Learning

Supervised learning:

- Make predictions, classify data, perform regression
- Dataset: $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)$

indoor

outdoor

Features / Covariates / Input

independent var.

Labels / Outputs

dependent variables.

- Goal: find function $f : X \to Y$ to predict label on **new** data

Back to Supervised Learning

prediction of lable   true lable

How do we know a function f is good?

- Intuitively: "matches" the dataset $f(x_i) \approx y_i$
- More concrete: pick a **loss function** to measure this: $\ell(f(x), y)$
- Training loss/empirical loss/empirical risk

Loss / Cost / Objective Function

(x_i, y_i)

$$\frac{1}{n}\sum_{i=1}^{n}\ell(f(x_i), y_i)$$

- Find a f that minimizes the loss on the training data (ERM)

"Empirical Risk Minimization"

Loss Functions

What should the loss look like?

- If $f(x_i) \approx y_i$, should be small (0 if equal!)
- For classification: 0/1 loss $\ell(f(x), y) = 1\{f(x_i) \neq y_i\}$
- For regression, square loss $\ell(f(x), y) = (f(x_i) - y_i)^2$

Others too! We'll see more.

logistic loss

$\ell(f(x), y) = \begin{cases} 1 & f(x) \neq y \\ 0 & f(x) = y \end{cases}$

$\ell(f(x), y) = |f(x_i) - y_i|$

$\ell(f(x), y) = |f(x_i) - y_i|^\alpha$

$\alpha > 0$

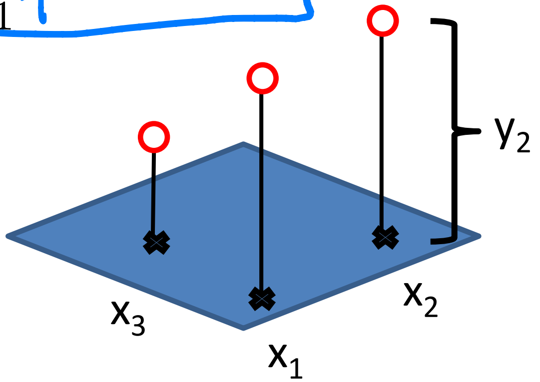
Functions/Models

$f(x_1) = y_1$
$f(x_2) = y_2$
$\vdots$
$f(x_n) = y_n$

$f(x) = 0$ for all other x.

The function f is usually called a model

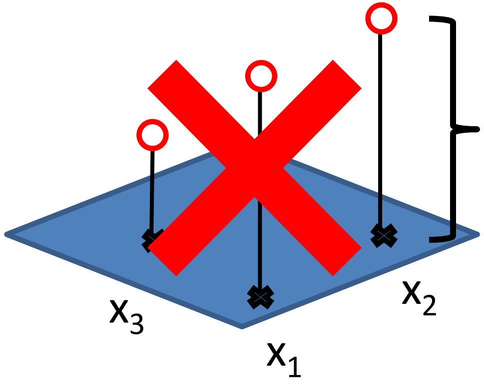
- Which possible functions should we consider?
- One option: **all functions**
  - Not a good choice. Consider $$f(x) = \sum_{i=1}^{n} 1\{x = x_i\} y_i$$
  - Training loss: **zero.** Can't do better!
  - How will it do on x not in the training set?



Functions/Models

Don't want all functions

- Instead, pick a specific class
- Parametrize it by weights/parameters
- **Example:** linear models



$$f(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_d x_d = \theta_0 + x^T \theta$$

Weights/ Parameters

$$\vec{\theta} = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_d \end{pmatrix} \quad \vec{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix}.$$

Training the Model

- Parametrize it by weights/parameters
- Minimize the loss

Best parameters = best function *f*

$$\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} \ell(f_{\theta}(x_i), y_i)$$

$$= \frac{1}{n} \sum_{i=1}^{n} \ell(\theta_0 + x_i^T \theta, y_i)$$

Linear model class *f*

$$= \frac{1}{n} \sum_{i=1}^{n} (\theta_0 + x_i^T \theta - y_i)^2$$

Square loss

How Do We Minimize?

- Need to solve something that looks like $\min_{\theta} g(\theta)$
  - Optimization problem; many algorithms
- **Gradient descent:**
  - start at some $\theta^{(0)}$
  - repeat till convergence:

terminate if $\|\nabla g(\theta^{(j)})\|_2 < 10^{-6}$ or after 1000 iteration

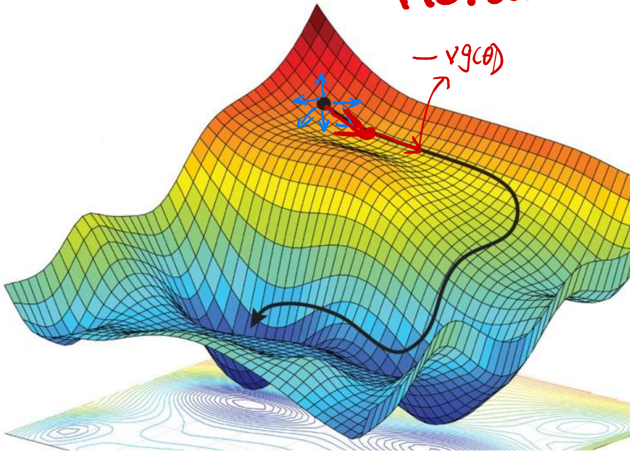
$$\theta^{(j+1)} = \theta^{(j)} - \gamma \nabla g(\theta^{(j)})$$

Next solution

Current solution

Learning Rate (a constant)

Gradient of the loss, evaluated at current sol.



M. Hutson

How Do We Minimize?

- Need to solve something that looks like $\min_{\theta \in \mathbb{R}^d} g(\theta)$
  - Optimization problem; many algorithms

$\nabla g(\theta) \in \mathbb{R}^d$

- **Gradient descent:**
  - start at some $\theta^{(0)}$
  - repeat till convergence:

$$\begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_d \end{pmatrix} \leftarrow \begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_d \end{pmatrix} - \eta \begin{pmatrix} (\nabla g(\theta))_1 \\ (\nabla g(\theta))_2 \\ \vdots \\ (\nabla g(\theta))_d \end{pmatrix}$$

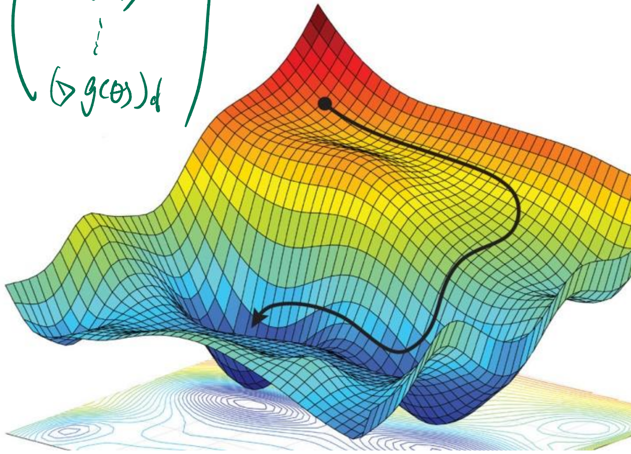
$$\theta^{(j+1)} = \theta^{(j)} - \gamma \nabla g(\theta^{(j)})$$

Gradient of the loss, evaluated at current sol.

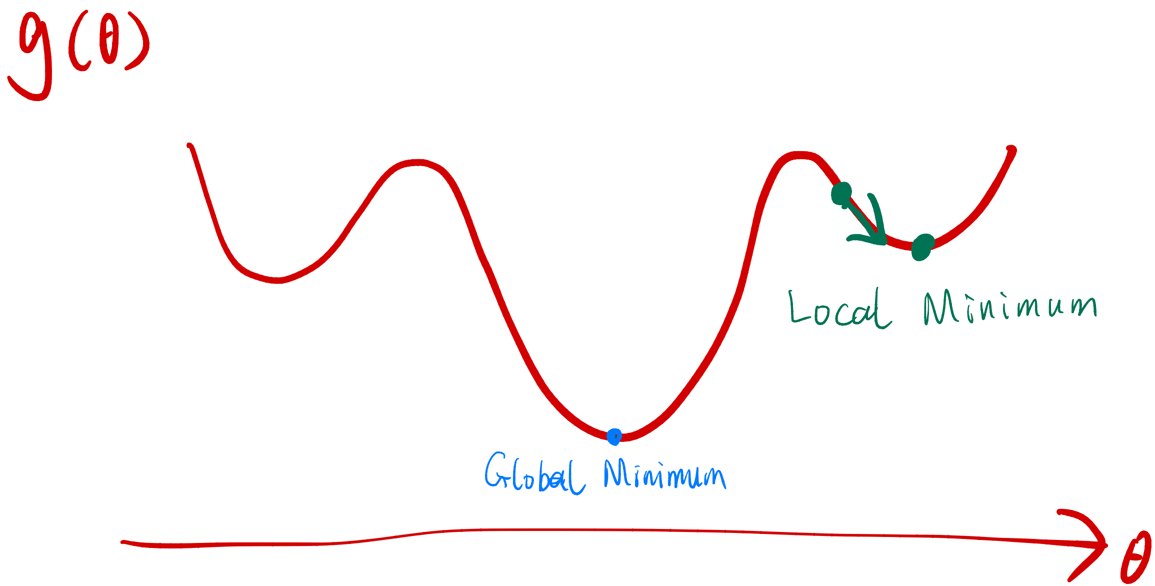
Next solution

Current solution

Learning Rate (a constant)



M. Hutson



g(θ)
Local Minimum
Global Minimum
θ

How Do We Minimize?

- Gradient descent
  - You’ll implement this in HW5
- Popular in practice: **stochastic gradient descent (SGD)**
- Most algorithms iterative:
  - find some sequence of points heading towards the optimum

Train vs Test

f generalizes ⟺ f performs well on points not training set

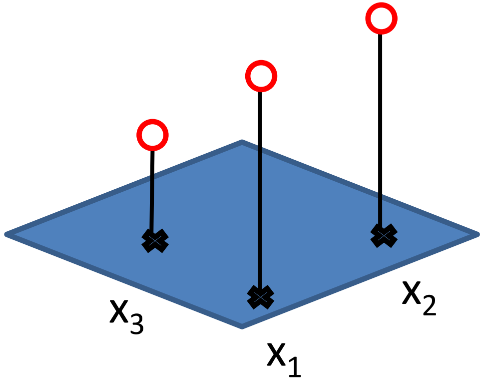
Now we've trained, have some *f* parametrized by θ

- Train loss is small → *f* predicts most x_i correctly
- How does *f* do on points not in training set? **"Generalizes!"**
- To evaluate this, use a **test** set. Do **not** train on it!

$$(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n) \quad (\mathbf{x}_{n+1}, y_{n+1}), \ldots, (\mathbf{x}_{n+p}, y_{n+p})$$

Training Data — Test Data

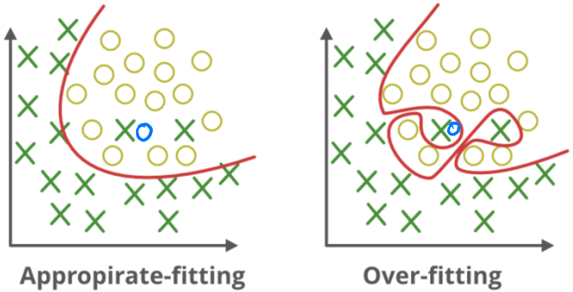
Train vs Test



Use the test set to evaluate *f*

- Why? Back to our “perfect” train function
- Training loss: 0. Every point matched perfectly
- How does it do on test set? **Fails completely**!

- Test set helps detect **overfitting**
  - Overfitting: too focused on train points
  - “Bigger” class: more prone to overfit
    - Need to consider **model capacity**



GFG

“Underfitting” = class inflexible, loss on training set big

Break & Quiz

Q 2.1: When we train a model, we are

- A. Optimizing the parameters and keeping the features fixed.
- B. Optimizing the features and keeping the parameters fixed.
- C. Optimizing the parameters and the features.
- D. Keeping parameters and features fixed and changing the predictions.

Break & Quiz

Q 2.1: When we train a model, we are

- **A. Optimizing the parameters and keeping the features fixed.**
- B. Optimizing the features and keeping the parameters fixed.
- C. Optimizing the parameters and the features.
- D. Keeping parameters and features fixed and changing the predictions.

Break & Quiz

Q 2.1: When we train a model, we are

- **A. Optimizing the parameters and keeping the features fixed.**
- B. Optimizing the features and keeping the parameters fixed) (Feature vectors x_i don't change during training).
- C. Optimizing the parameters and the features. (Same as B)
- D. Keeping parameters and features fixed and changing the predictions. (We can't train if we don't change the parameters)

Break & Quiz

- **Q 2.2**: You have trained a classifier, and you find there is significantly **higher** loss on the test set than the training set. What is likely the case?

- A. You have accidentally trained your classifier on the test set.
- B. Your classifier is generalizing well.
- C. Your classifier is generalizing poorly.
- D. Your classifier is ready for use.

train a simpler model.

Break & Quiz

- **Q 2.2**: You have trained a classifier, and you find there is significantly **higher** loss on the test set than the training set. What is likely the case?

- A. You have accidentally trained your classifier on the test set.
- B. Your classifier is generalizing well.
- **C. Your classifier is generalizing poorly.**
- D. Your classifier is ready for use.

Break & Quiz

- **Q 2.2**: You have trained a classifier, and you find there is significantly **higher** loss on the test set than the training set. What is likely the case?

- A. You have accidentally trained your classifier on the test set. **(No, this would make test loss lower)**
- B. Your classifier is generalizing well. **(No, test loss is high means poor generalization)**
- **C. Your classifier is generalizing poorly.**
- D. Your classifier is ready for use. **(No, will perform poorly on new data)**

Break & Quiz

- **Q 2.3**: You have trained a classifier, and you find there is significantly **lower** loss on the test set than the training set. What is likely the case?

- A. You have accidentally trained your classifier on the test set.
- B. Your classifier is generalizing well.
- C. Your classifier is generalizing poorly.
- D. Your classifier needs further training.

Break & Quiz

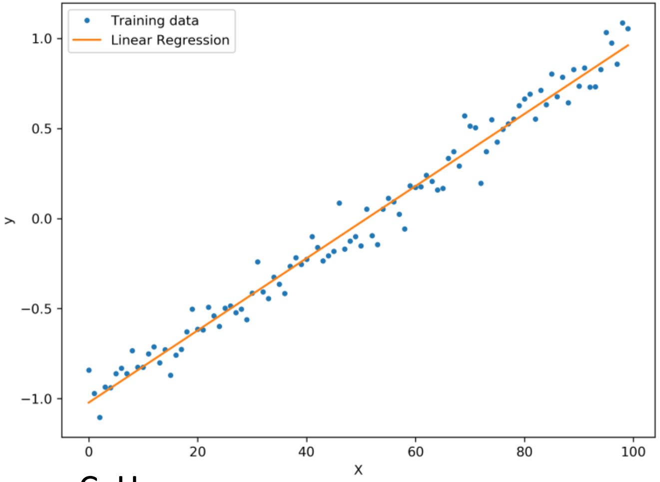
- **Q 2.3**: You have trained a classifier, and you find there is significantly **lower** loss on the test set than the training set. What is likely the case?

- **A. You have accidentally trained your classifier on the test set.** (Loss will usually be the lowest on the data set on which a model has been trained)
- B. Your classifier is generalizing well.
- C. Your classifier is generalizing poorly.
- D. Your classifier needs further training.

Linear Regression

Simplest type of regression problem.

- **Inputs**: $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)$ $\in \mathbb{R}$
  - x's are vectors, y's are scalars.
  - "**Linear**": predict a linear combination of x components + intercept



C. Hansen

$$f(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_d x_d = \theta_0 + x^T \theta$$

- **Want**: parameters θ_0, θ

Linear Regression Setup

- Goal: figure out how to minimize square loss
- Train set $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)$
  - Model $f(x) = \theta_0 + x^T\theta$, wrap intercept: $f(x) = x^T\theta$
  - Take train data and make it a matrix

$$X \in \mathbb{R}^{n \times d}$$

  - Then, square loss is

$$\frac{1}{n}\sum_{i=1}^{n}(x_i^T\theta - y_i)^2 = \frac{1}{n}\|X\theta - y\|^2$$

Finding The Optimal Parameters

Have our loss: $g(\theta) = \frac{1}{n}\|X\theta - y\|^2$ ← least squares problem

- Could optimize it with GD, SGD, etc…
- Explicit formula for the minimum

$$\nabla g(\theta) = \frac{2}{n} X^T(X\theta - y)$$

Hat: indicates an estimate

$$\hat{\theta} = (X^TX)^{-1}X^Ty$$

$$\nabla g(\theta) = \begin{pmatrix} \frac{\partial g(\theta)}{\partial \theta_1} \\ \vdots \\ \frac{\partial g(\theta)}{\partial \theta_d} \end{pmatrix}$$

- Why use GD/SGD when we have explicit formula?
  - computing inverse can be expensive.
  - X^TX may be not invertible. (there are multiple optimal solution)

How Good are the Optimal Parameters?

Now we have parameters $\hat{\theta} = (X^T X)^{-1} X^T y$

- How good are they?
- Predictions are $f(x_i) = \hat{\theta}^T x_i = ((X^T X)^{-1} X^T y)^T x_i$
- Errors ("**residuals**") on training set

$$|y_i - f(x_i)| = |y_i - \hat{\theta}^T x_i| = |y_i - ((X^T X)^{-1} X^T y)^T x_i|$$

- Small residuals: fit **training** set well
- May want to use a **test** set to check

class i | Not class i.

Linear Regression → Classification?

class i | class j

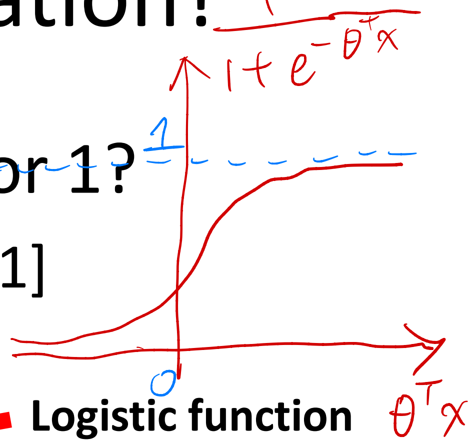
What if we want the same idea, but y is 0 or 1?

- Need to convert the $\theta^T x$ to a probability in [0,1]

f(x) =

$$p(y=1|x) = \frac{1}{1+\exp(-\theta^T x)}$$

← **Logistic function**



Why does this work?

- If $\theta^T x$ is really big, $\exp(-\theta^T x)$ is really small → *p* close to 1
- If really negative exp is huge → *p* close to 0

"Logistic Regression"

Linear Regression → Classification?

What if we want the same idea, but y is 0 or 1?

- Need to convert the $\theta^T x$ to a probability in [0,1]

$$p(y=1|x) = \frac{1}{1+\exp(-\theta^T x)}$$

Logistic function

Why does this work?

- If $\theta^T x$ is really big, $\exp(-\theta^T x)$ is really small → *p* close to 1
- If really negative exp is huge → *p* close to 0

"Logistic Regression"