10-601 Introduction to Machine Learning

Machine Learning Department
School of Computer Science
Carnegie Mellon University

Logistic Regression

Matt Gormley
Lecture 8
Feb. 12, 2018

10-601 Introduction to Machine Learning

Machine Learning Department
School of Computer Science
Carnegie Mellon University

~~Logistic Regression~~
Probabilistic Learning

Matt Gormley
Lecture 8
Feb. 12, 2018

Reminders

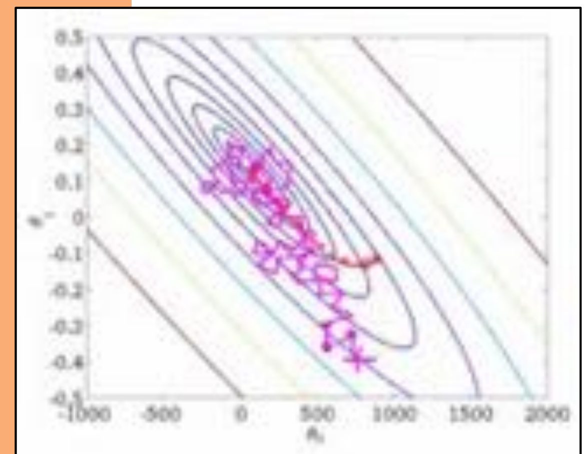
- **Homework 3: KNN, Perceptron, Lin.Reg.**
  - Out: Wed, Feb 7
  - Due: Wed, Feb 14 at 11:59pm

STOCHASTIC GRADIENT DESCENT

Stochastic Gradient Descent (SGD)

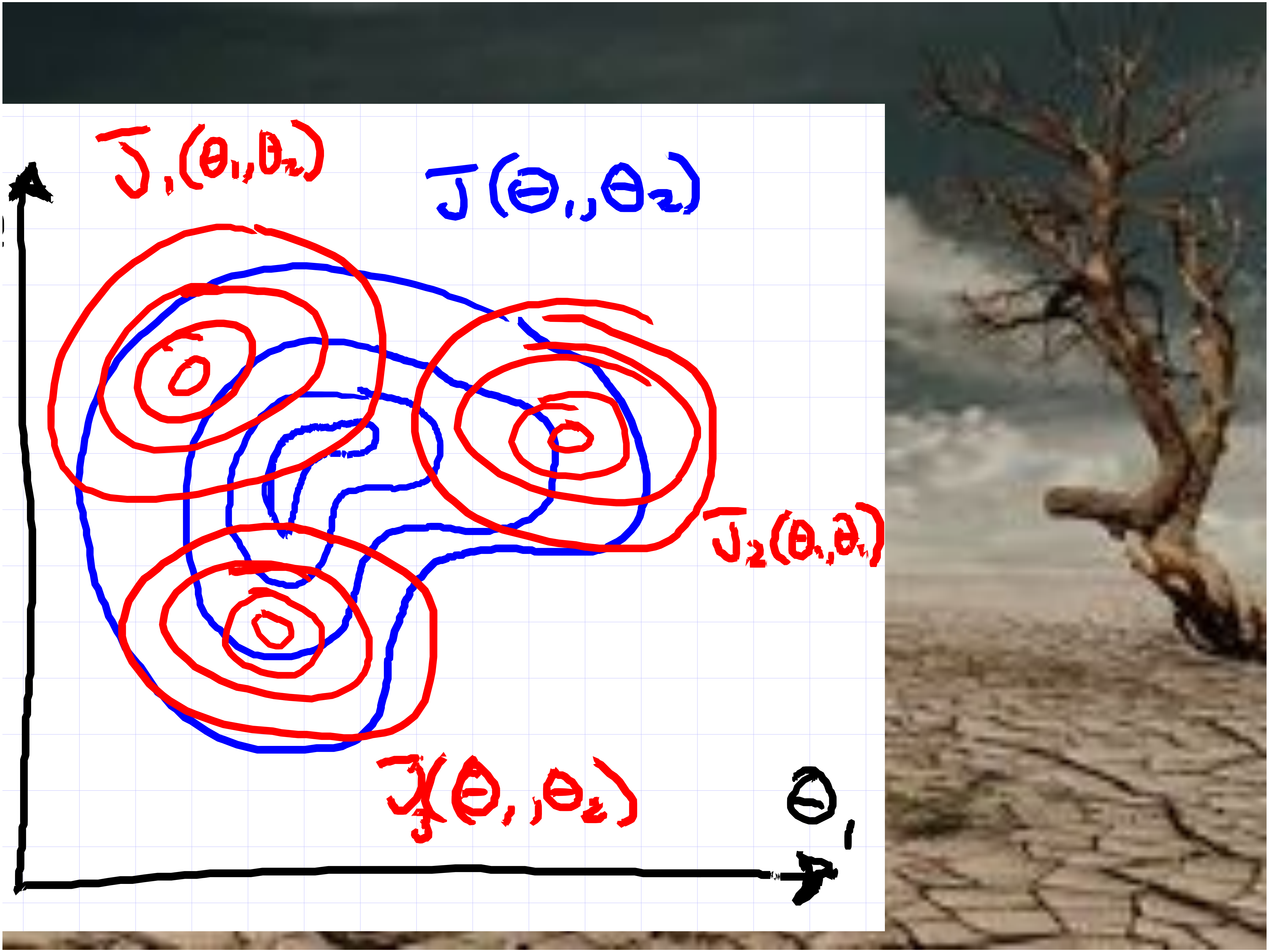
Algorithm 2 Stochastic Gradient Descent (SGD)

procedure $\text{SGD}(\mathcal{D}, \boldsymbol{\theta}^{(0)})$
  $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta}^{(0)}$
  while not converged **do**
    for $i \in \text{shuffle}(\{1, 2, \ldots, N\})$ **do**
      $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \lambda \nabla_{\boldsymbol{\theta}} J^{(i)}(\boldsymbol{\theta})$
  return $\boldsymbol{\theta}$



We need a per-example objective:

$$\text{Let } J(\boldsymbol{\theta}) = \sum_{i=1}^{N} J^{(i)}(\boldsymbol{\theta})$$


$J_1(\theta_1,\theta_2)$
$J(\theta_1,\theta_2)$
$J_2(\theta_1,\theta_2)$
$J_3(\theta_1,\theta_2)$
θ_1

Expectations of Gradients

$$\frac{dJ(\vec{\theta})}{d\theta_j} = \frac{1}{N}\sum_{i=1}^{N} \frac{d}{d\theta_j}\left(J_i(\vec{\theta})\right)$$

$$\nabla J(\vec{\theta}) = \begin{array}{c} \\ j^{th} \\ \\ \end{array} \begin{bmatrix} \vdots \\ \_\_\_ \\ \vdots \end{bmatrix} = \frac{1}{N}\sum_{i=1}^{N} \nabla J_i(\vec{\theta})$$

Recall: for any discrete r.v. X

$$E_X[f(X)] \triangleq \sum_x P(X=x) f(x)$$

Q: What is the expected value of a randomly chosen $\nabla J_i(\theta)$?

Let $I \sim \text{Uniform}(\{1, \ldots, N\})$

$\Rightarrow P(I=i) = \frac{1}{N}$ if $i \in \{1 \ldots N\}$

$$E_I[\nabla J_I(\vec{\theta})] = \sum_{i=1}^{N} P(I=i) \nabla J_i(\vec{\theta})$$

$$= \frac{1}{N}\sum_{i=1}^{N} \nabla J_i(\vec{\theta})$$

$$= \nabla J(\vec{\theta})$$

Convergence of Optimizers

Convergence Analysis:

true unknown min

Def: convergence is when $J(\vec{\theta}) - J(\vec{\theta}^*) < \epsilon$

Methods	Steps to Converge	Computation per iteration
Newton's Method	$O(\ln \ln 1/\epsilon)$	$\nabla J(\theta)$ $\nabla^2 J(\theta)$ $\leftarrow O(NM^2)$
GD	$O(\ln 1/\epsilon)$	$\nabla J(\vec{\theta})$ $\leftarrow O(NM)$
SGD	$O(1/\epsilon)$	$\nabla J_i(\theta)$ $\leftarrow O(M)$

not covered

"almost sure" convergence

lots of caveats and conditions

way less computation

Takeaway: SGD has much slower asymptotic convergence. but is often faster in practice.

Optimization Objectives

You should be able to…

- Apply gradient descent to optimize a function
- Apply stochastic gradient descent (SGD) to optimize a function
- Apply knowledge of zero derivatives to identify a closed-form solution (if one exists) to an optimization problem
- Distinguish between convex, concave, and nonconvex functions
- Obtain the gradient (and Hessian) of a (twice) differentiable function

Linear Regression Objectives

You should be able to…

- Design k-NN Regression and Decision Tree Regression
- Implement learning for Linear Regression using three optimization techniques: (1) closed form, (2) gradient descent, (3) stochastic gradient descent
- Choose a Linear Regression optimization technique that is appropriate for a particular dataset by analyzing the tradeoff of computational complexity vs. convergence speed
- Distinguish the three sources of error identified by the bias-variance decomposition: bias, variance, and irreducible error.

PROBABILISTIC LEARNING

Probabilistic Learning

Function Approximation

Previously, we assumed that our output was generated using a **deterministic target function:**

$$\mathbf{x}^{(i)} \sim p^*(\cdot)$$

$$y^{(i)} = c^*(\mathbf{x}^{(i)})$$

Our goal was to learn a hypothesis h(**x**) that best approximates c^*(**x**)

Probabilistic Learning

Today, we assume that our output is **sampled** from a conditional **probability distribution:**

$$\mathbf{x}^{(i)} \sim p^*(\cdot)$$

$$y^{(i)} \sim p^*(\cdot|\mathbf{x}^{(i)})$$

Our goal is to learn a probability distribution p(y|**x**) that best approximates p^*(y|**x**)

Robotic Farming

	Deterministic	**Probabilistic**
Classification (binary output)	Is this a picture of a wheat kernel?	Is this plant drought resistant?
Regression (continuous output)	How many wheat kernels are in this picture?	What will the yield of this plant be?

Maximum Likelihood Estimation

The principle of Maximum likelihood estimator (MLE):

Choose parameters that make the data "most likely".

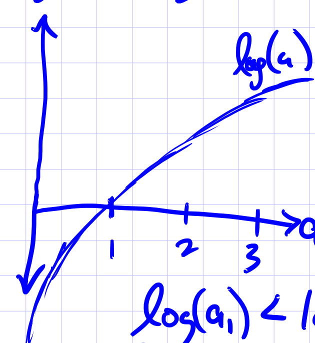
Assumptions: Data generated iid from distribution $p^*(x \mid \vec{\theta}^*)$ and comes from a family of distr. parameterized $\theta \in \Theta$ ← set of possible parameters

Formally:

$$\theta_{MLE} = \underset{\theta \in \Theta}{\arg\max}\; p(D \mid \theta)$$

$$= \underset{\theta \in \Theta}{\arg\max}\; \log p(D \mid \theta)$$

since log is monotonic



$\log(a_1) < \log(a_2)$ iff $a_1 < a_2$

$\Rightarrow \log(f(a_1)) < \log(f(a_2))$ iff $f(a_1) < f(a_2)$

usually a continuous optimization →

$$= \underset{\theta \in \Theta}{\arg\max}\; \ell(\theta)$$

where $\ell(\theta) \triangleq \log p(D \mid \theta)$

"log-likelihood"

treat as function of θ where D is constant

Learning from Data (Frequentist)

Whiteboard

- Principle of Maximum Likelihood Estimation (MLE)
- Strawmen:
  - Example: Bernoulli
  - Example: Gaussian
  - Example: Conditional #1
    (Bernoulli conditioned on Gaussian)
  - Example: Conditional #2
    (Gaussians conditioned on Bernoulli)

Outline

- **Motivation:**
  - Choosing the right classifier
  - Example: Image Classification
- **Logistic Regression**
  - Background: Hyperplanes
  - Data, Model, Learning, Prediction
  - Log-odds
  - Bernoulli interpretation
  - Maximum Conditional Likelihood Estimation
- **Gradient descent for Logistic Regression**
  - Stochastic Gradient Descent (SGD)
  - Computing the gradient
  - Details (learning rate, finite differences)

MOTIVATION:
LOGISTIC REGRESSION

Example: Image Classification

- ImageNet LSVRC-2010 contest:
  - **Dataset:** 1.2 million labeled images, 1000 classes
  - **Task**: Given a new image, label it with the correct class
  - **Multiclass** classification problem
- Examples from http://image-net.org/


IMAGENET
Bird


IMAGENET
German iris, Iris kochii


IMAGENET
Court, courtyard

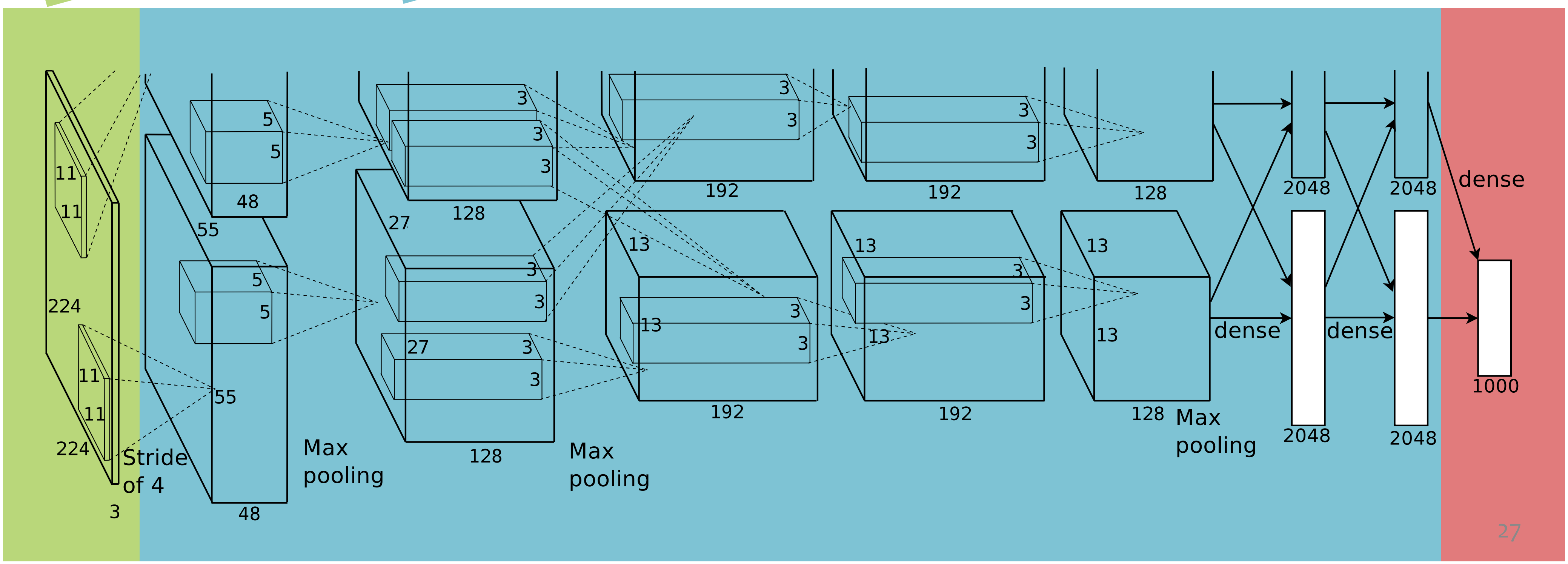
Example: Image Classification

CNN for Image Classification
(Krizhevsky, Sutskever & Hinton, 2011)
17.5% error on ImageNet LSVRC-2010 contest

Input image (pixels)

- Five convolutional layers (w/max-pooling)
- Three fully connected layers

1000-way softmax



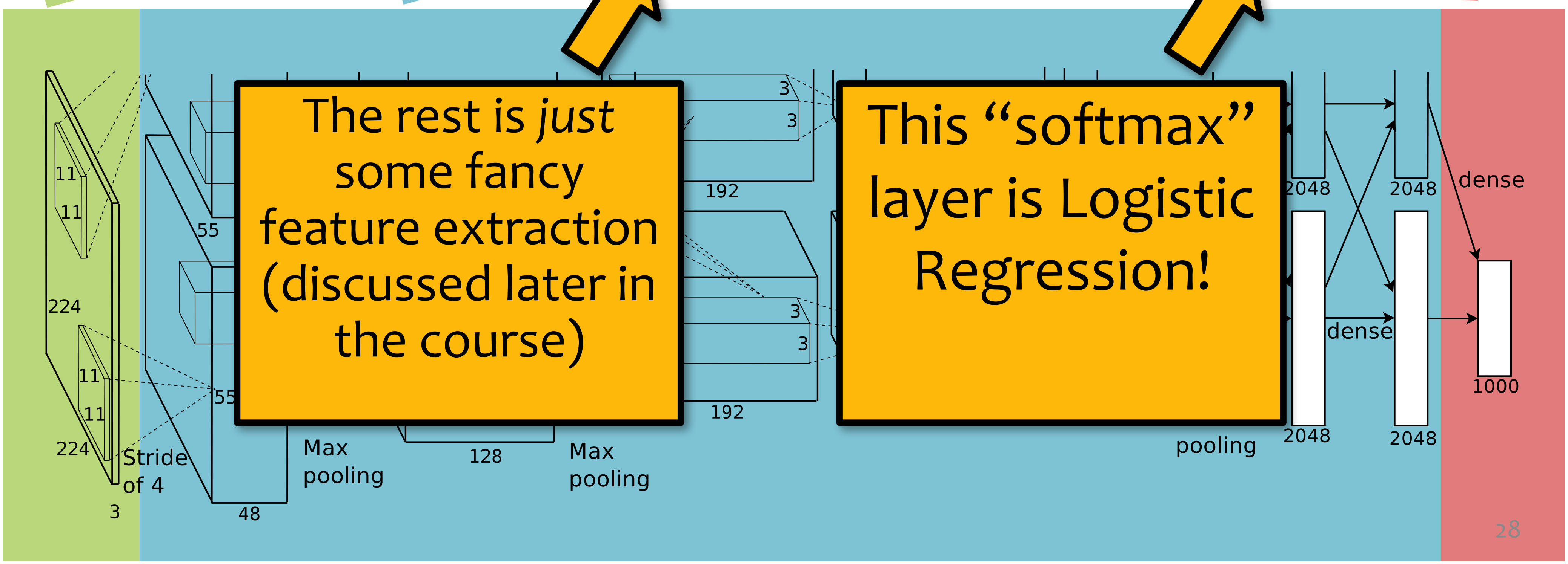
Example: Image Classification

CNN for Image Classification
(Krizhevsky, Sutskever & Hinton, 2011)
17.5% error on ImageNet LSVRC-2010 contest

Input image (pixels)

- Five convolutional layers (w/max-pooling)
- Three fully connected layers

1000-way softmax



The rest is *just* some fancy feature extraction (discussed later in the course)

This "softmax" layer is Logistic Regression!

LOGISTIC REGRESSION

Logistic Regression

Data: Inputs are continuous vectors of length K. Outputs are discrete.

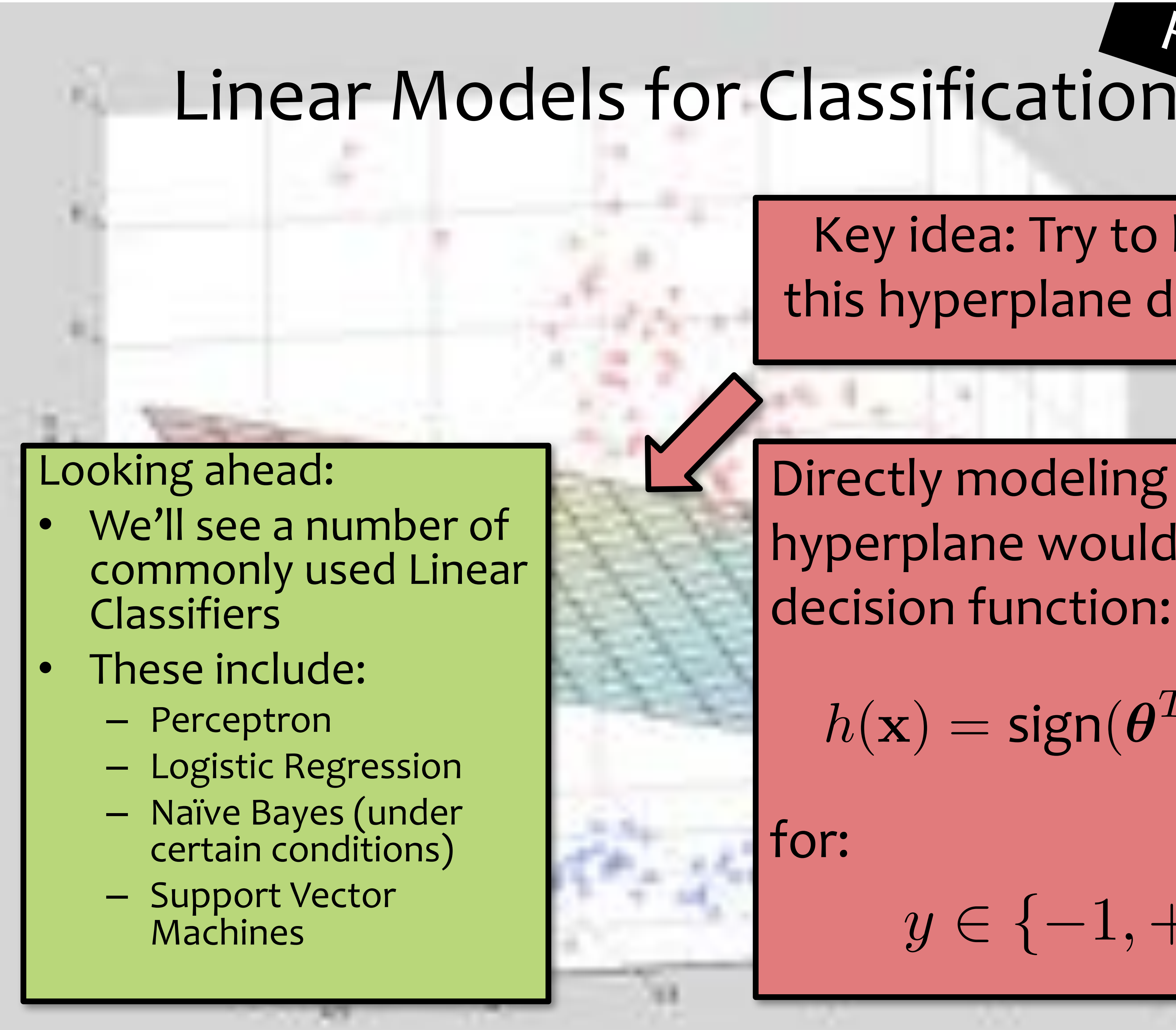
$$\mathcal{D} = \{\mathbf{x}^{(i)}, y^{(i)}\}_{i=1}^{N} \text{ where } \mathbf{x} \in \mathbb{R}^M \text{ and } y \in \{0, 1\}$$

We are back to classification.

Despite the name logistic **regression.**

Linear Models for Classification

Recall…



Key idea: Try to learn this hyperplane directly

Directly modeling the hyperplane would use a decision function:

$$h(\mathbf{x}) = \text{sign}(\boldsymbol{\theta}^T \mathbf{x})$$

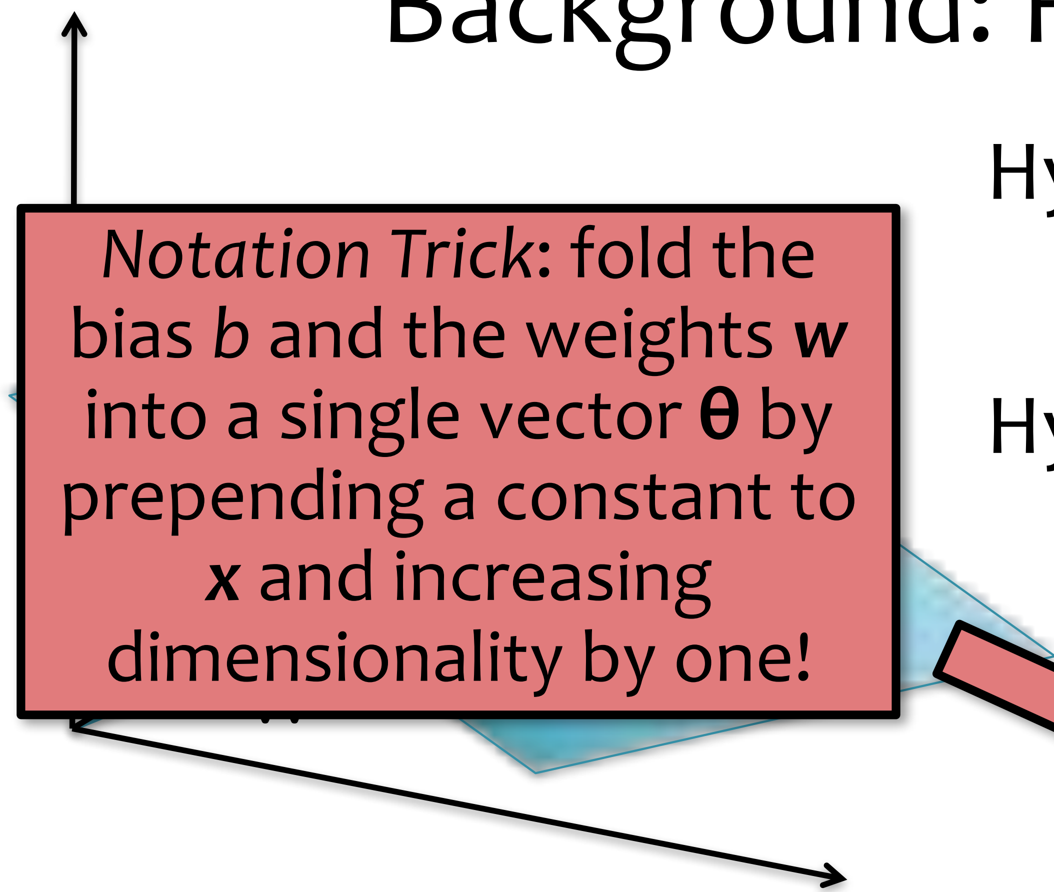
for:

$$y \in \{-1, +1\}$$

Looking ahead:

- We'll see a number of commonly used Linear Classifiers
- These include:
  - Perceptron
  - Logistic Regression
  - Naïve Bayes (under certain conditions)
  - Support Vector Machines

Background: Hyperplanes



Notation Trick: fold the bias b and the weights $\mathbf{w}$ into a single vector $\boldsymbol{\theta}$ by prepending a constant to $\mathbf{x}$ and increasing dimensionality by one!

Hyperplane (Definition 1):

$$\mathcal{H} = \{\mathbf{x} : \mathbf{w}^T \mathbf{x} = b\}$$

Hyperplane (Definition 2):

$$\mathcal{H} = \{\mathbf{x} : \boldsymbol{\theta}^T \mathbf{x} = 0 \text{ and } x_0 = 1\}$$

$$\boldsymbol{\theta} = [b, w_1, \ldots, w_M]^T$$

Half-spaces:

$$\mathcal{H}^+ = \{\mathbf{x} : \boldsymbol{\theta}^T \mathbf{x} > 0 \text{ and } x_0 = 1\}$$

$$\mathcal{H}^- = \{\mathbf{x} : \boldsymbol{\theta}^T \mathbf{x} < 0 \text{ and } x_0 = 1\}$$

Using gradient ascent for linear classifiers

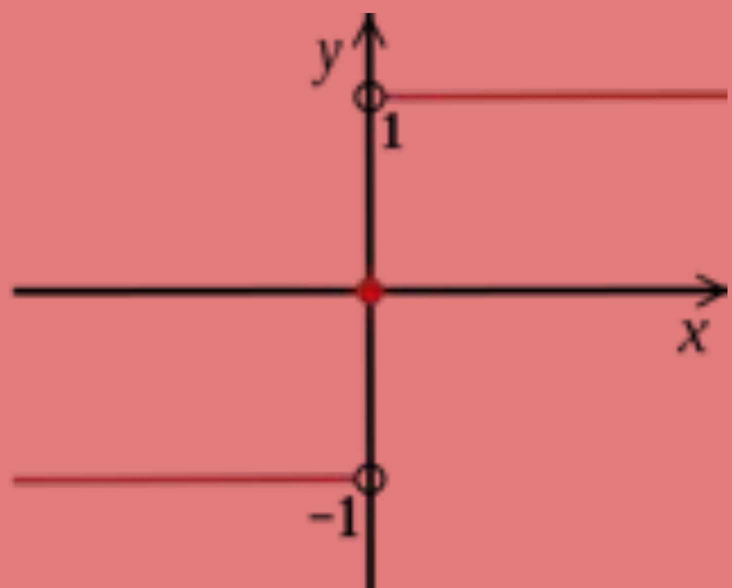
Key idea behind today's lecture:

1. Define a linear classifier (logistic regression)
2. Define an objective function (likelihood)
3. Optimize it with gradient descent to learn parameters
4. Predict the class with highest probability under the model

Using gradient ascent for linear classifiers

This decision function isn't differentiable:

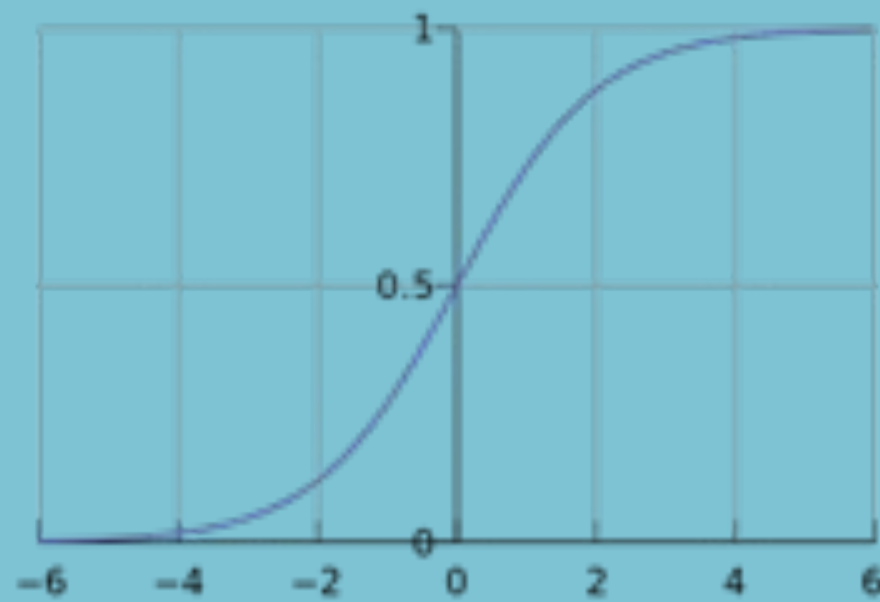
$$h(\mathbf{x}) = \text{sign}(\boldsymbol{\theta}^T \mathbf{x})$$



sign(x)

Use a differentiable function instead:

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

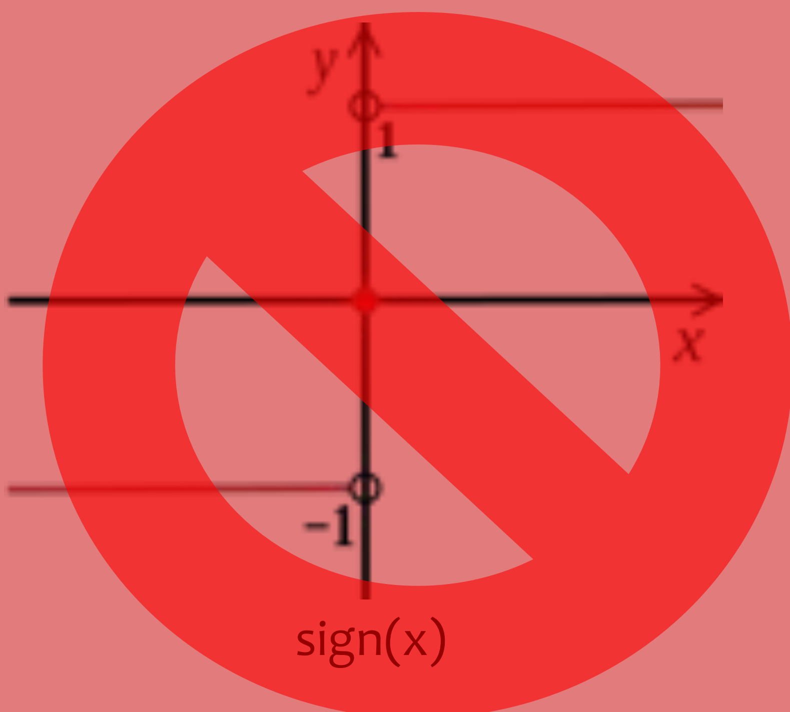


$$\text{logistic}(u) \equiv \frac{1}{1 + e^{-u}}$$

Using gradient ascent for linear classifiers

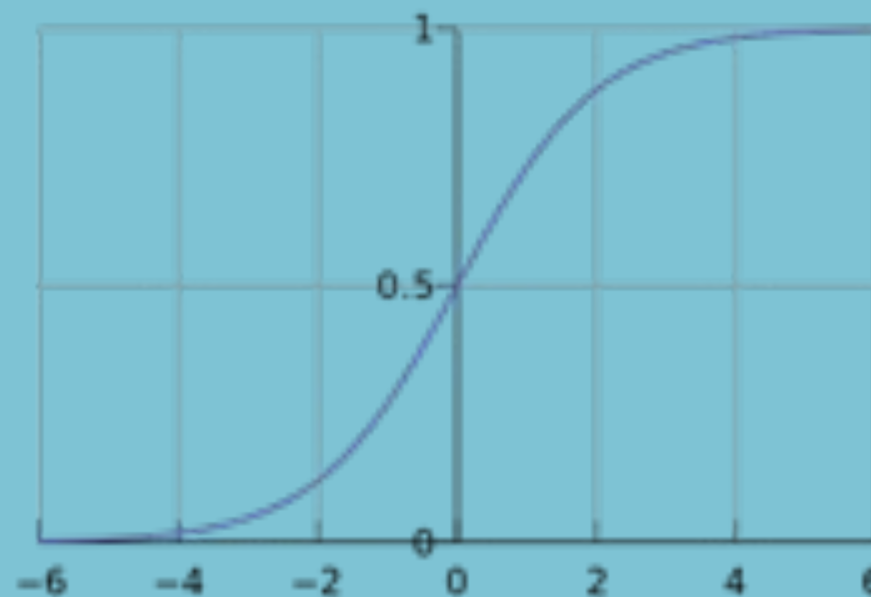
This decision function isn't differentiable:

$$h(\mathbf{x}) = \text{sign}(\boldsymbol{\theta}^T \mathbf{x})$$



Use a differentiable function instead:

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



$$\text{logistic}(u) \equiv \frac{1}{1 + e^{-u}}$$

Logistic Regression

Data: Inputs are continuous vectors of length K. Outputs are discrete.

$$\mathcal{D} = \{\mathbf{x}^{(i)}, y^{(i)}\}_{i=1}^N \text{ where } \mathbf{x} \in \mathbb{R}^M \text{ and } y \in \{0, 1\}$$

Model: Logistic function applied to dot product of parameters with input vector.

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

Learning: finds the parameters that minimize some objective function.

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\, J(\boldsymbol{\theta})$$

Prediction: Output is the most probable class.

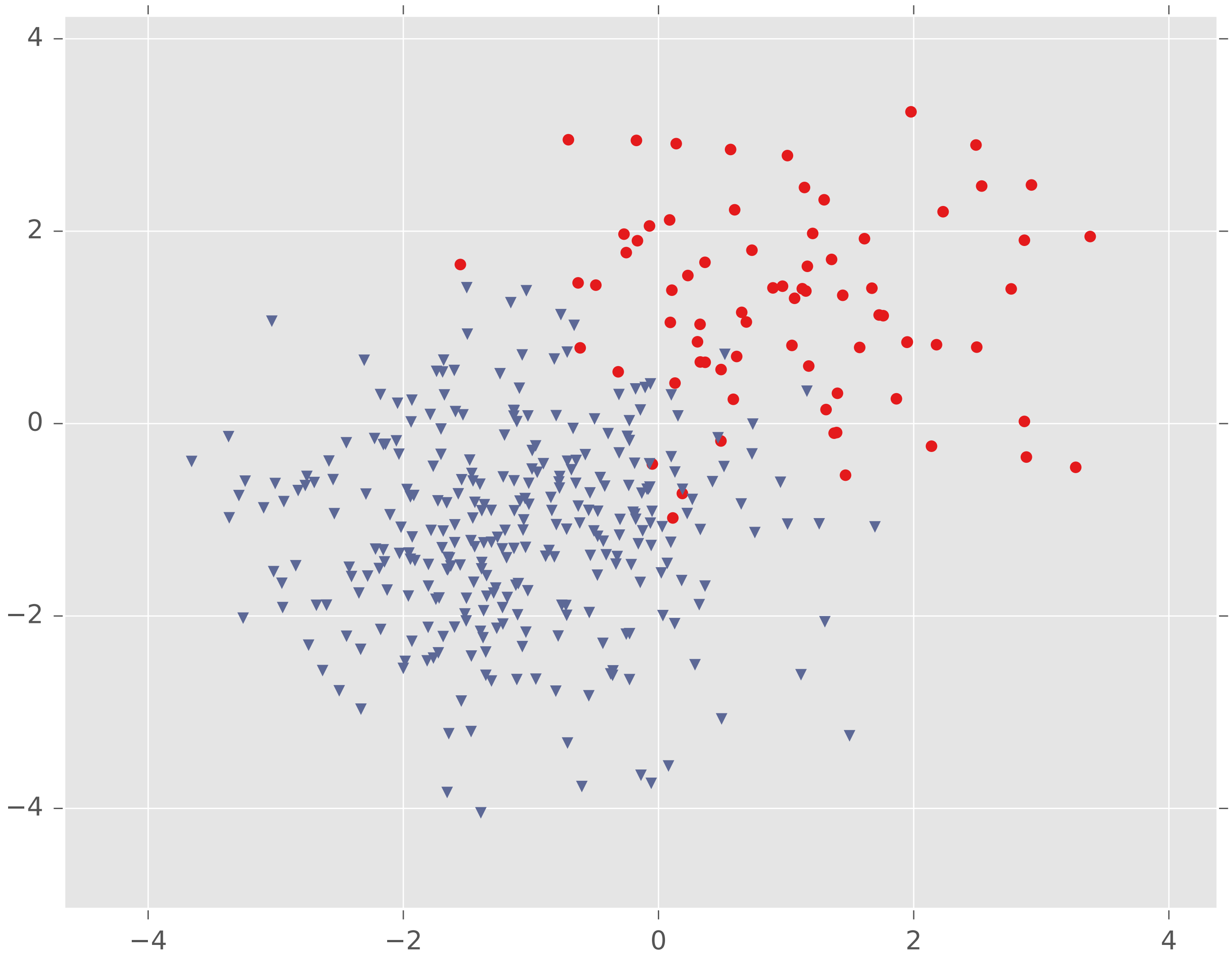
$$\hat{y} = \underset{y \in \{0,1\}}{\operatorname{argmax}}\, p_{\boldsymbol{\theta}}(y|\mathbf{x})$$

Logistic Regression

Whiteboard

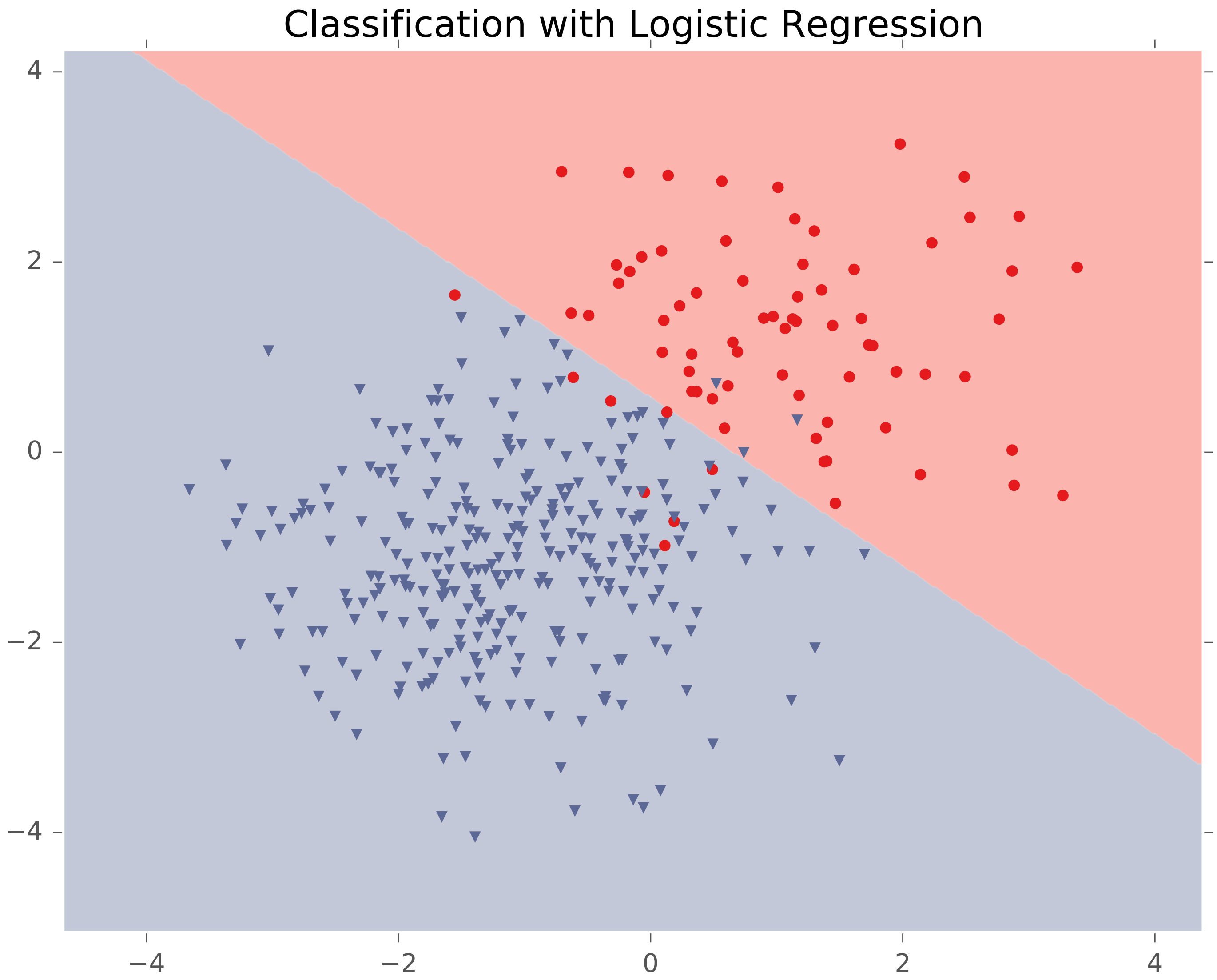
- Bernoulli interpretation
- Logistic Regression Model
- Decision boundary

Logistic Regression



Logistic Regression

Logistic Regression



LEARNING LOGISTIC REGRESSION

Maximum **Conditional** Likelihood Estimation

Learning: finds the parameters that minimize some objective function.

$$\boldsymbol{\theta}^* = \operatorname*{argmin}_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$$

We minimize the *negative* log conditional likelihood:

$$J(\boldsymbol{\theta}) = -\log \prod_{i=1}^{N} p_{\boldsymbol{\theta}}(y^{(i)} | \mathbf{x}^{(i)})$$

Why?

1. We can't maximize likelihood (as in Naïve Bayes) because we don't have a joint model p(x,y)
2. It worked well for Linear Regression (least squares is MCLE)

Maximum **Conditional** Likelihood Estimation

Learning: Four approaches to solving $\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\, J(\boldsymbol{\theta})$

Approach 1: Gradient Descent
(take larger – more certain – steps opposite the gradient)

Approach 2: Stochastic Gradient Descent (SGD)
(take many small steps opposite the gradient)

Approach 3: Newton's Method
(use second derivatives to better follow curvature)

Approach 4: Closed Form???
(set derivatives equal to zero and solve for parameters)

Maximum **Conditional** Likelihood Estimation

Learning: Four approaches to solving $\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\, J(\boldsymbol{\theta})$

Approach 1: Gradient Descent
(take larger – more certain – steps opposite the gradient)

Approach 2: Stochastic Gradient Descent (SGD)
(take many small steps opposite the gradient)

Approach 3: Newton's Method
(use second derivatives to better follow curvature)

~~**Approach 4:** Closed Form???~~
~~(set derivatives equal to zero and solve for parameters)~~

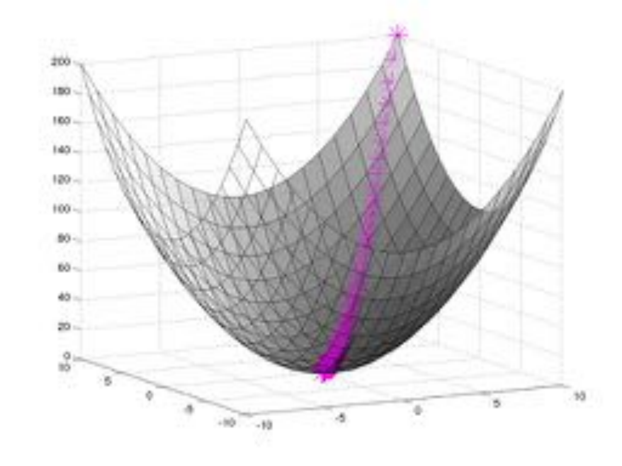
Logistic Regression does not have a closed form solution for MLE parameters.

Recall…

Gradient Descent

Algorithm 1 Gradient Descent

procedure GD($\mathcal{D}$, $\boldsymbol{\theta}^{(0)}$)
  $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta}^{(0)}$
  while not converged **do**
    $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \lambda \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$
  return $\boldsymbol{\theta}$



In order to apply GD to Logistic Regression all we need is the **gradient** of the objective function (i.e. vector of partial derivatives).

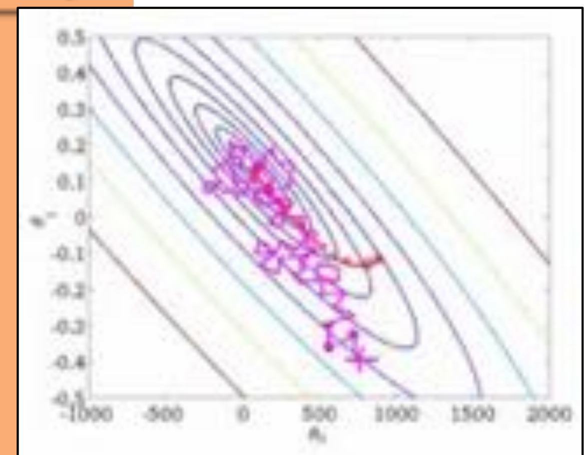
$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \begin{bmatrix} \frac{d}{d\theta_1} J(\boldsymbol{\theta}) \\ \frac{d}{d\theta_2} J(\boldsymbol{\theta}) \\ \vdots \\ \frac{d}{d\theta_M} J(\boldsymbol{\theta}) \end{bmatrix}$$

Stochastic Gradient Descent (SGD)

Recall…

Algorithm 1 Stochastic Gradient Descent (SGD)

procedure SGD($\mathcal{D}$, $\boldsymbol{\theta}^{(0)}$)
    $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta}^{(0)}$
    while not converged **do**
        for $i \in \text{shuffle}(\{1, 2, \ldots, N\})$ **do**
            $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \lambda \nabla_{\boldsymbol{\theta}} J^{(i)}(\boldsymbol{\theta})$
    return $\boldsymbol{\theta}$



We can also apply SGD to solve the MCLE problem for Logistic Regression.

We need a per-example objective:

$$\text{Let } J(\boldsymbol{\theta}) = \textstyle\sum_{i=1}^{N} J^{(i)}(\boldsymbol{\theta})$$
$$\text{where } J^{(i)}(\boldsymbol{\theta}) = -\log p_{\boldsymbol{\theta}}(y^i | \mathbf{x}^i).$$

GRADIENT FOR LOGISTIC REGRESSION

Learning for Logistic Regression

Whiteboard

– Partial derivative for Logistic Regression
– Gradient for Logistic Regression

Details: Picking learning rate

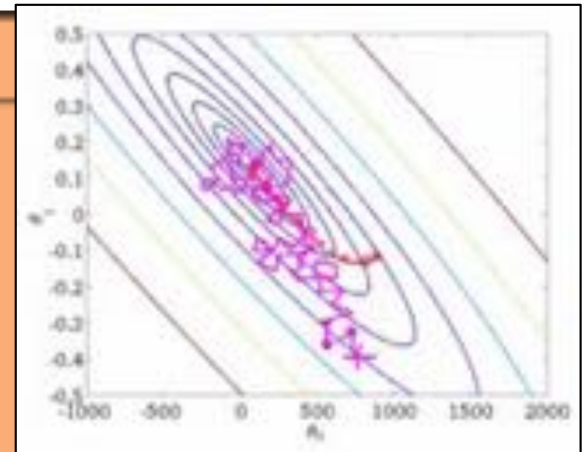
- Use grid-search in log-space over small values on a tuning set:
  - e.g., 0.01, 0.001, …
- Sometimes, decrease after each pass:
  - e.g factor of $1/(1 + dt)$, t=epoch
  - sometimes $1/t^2$
- Fancier techniques I won't talk about:
  - Adaptive gradient: scale gradient differently for each dimension (Adagrad, ADAM, ….)

Slide courtesy of William Cohen

SGD for Logistic Regression

Algorithm 1 SGD for Logistic Regression

procedure SGD($\mathcal{D}, \boldsymbol{\theta}^{(0)}$)
  $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta}^{(0)}$
  while not converged **do**
    for $i \in$ shuffle($\{1, 2, \ldots, N\}$) **do**
      $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \lambda(y^{(i)} - \rho^{(i)})\mathbf{x}^{(i)}$
      where $\rho^{(i)} := 1/(1 + \exp(-\boldsymbol{\theta}^T \mathbf{x}))$
  return $\boldsymbol{\theta}$



We can also apply SGD to solve the MCLE problem for Logistic Regression.

We need a per-example objective:

$$\text{Let } J(\boldsymbol{\theta}) = \sum_{i=1}^{N} J^{(i)}(\boldsymbol{\theta})$$
$$\text{where } J^{(i)}(\boldsymbol{\theta}) = -\log p_{\boldsymbol{\theta}}(y^i|\mathbf{x}^i).$$

Summary

1. Discriminative classifiers directly model the **conditional**, p(y|x)
2. Logistic regression is a **simple linear classifier**, that retains a **probabilistic** semantics
3. Parameters in LR are learned by **iterative optimization** (e.g. SGD)

Probabilistic Interpretation of Linear Regression

Whiteboard

- Conditional Likelihood
- Case #1: 1D Linear Regression
- Case #2: Multiple Linear Regression
- Equivalence: Predictions
- Equivalence: Learning