# Introduction to Machine Learning

Benjamin Peherstorfer

Fall 2020

# Today

- Last time
  - Least-squares regression
  - Recap concepts from probability theory (https://see.stanford.edu/materials/aimlcs229/cs229-prob.pdf)
  - Recommended reading: Recap concepts from linear algebra: https://see.stanford.edu/materials/aimlcs229/cs229-linalg.pdf

- Today
  - Finish up probability theory recap
  - Probabilistic interpretation of regression/classification
  - Reading: https://see.stanford.edu/materials/aimlcs229/cs229-notes1.pdf

- Announcements
  - Coming up: **blended** lab session (Wed, 9/16)

# Recap: define hypothesis space, define loss, then optimize

- Hypothesis space

$$\mathscr{H} = \{h_\theta : h_\theta(\boldsymbol{x}) = \sum_{i=0}^{n} \theta_i x_i = \boldsymbol{\theta}^T \boldsymbol{x}\}$$

- Loss function

$$J(\boldsymbol{\theta}) = \frac{1}{2} \sum_{i=1}^{N} \left(h_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}) - \boldsymbol{y}^{(i)}\right)^2$$

- Finding $h^*$ becomes an optimization problem

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$$

# Probabilistic point of view

- Consider probabilistic procedure that has generated data

- Identify the parameters that assign the highest probability to data that were observed

- Consider data generated from a Gaussian distribution

$$\{y^{(i)}\}_{i=1}^{N}$$

  - All data points are independently and identically distributed (iid)
  - The Gaussian distribution has unknown mean $\mu$ and variance $\sigma^2$
  - Write as probability

$$y^{(i)} \mid \mu, \sigma^2 \sim \mathcal{N}(\mu, \sigma^2)$$

# Probabilistic point of view (cont'd)

- Probability density function

$$P(y^{(i)} \mid \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(y^{(i)} - \mu)^2\right)$$

- We have $N$ **iid** data points $\{y^{(i)}\}_{i=1}^N$ with distribution

$$P(\{y^{(i)}\}_{i=1}^N \mid \mu, \sigma^2) = \prod_{i=1}^N \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(y^{(i)} - \mu)^2\right)$$

- Intuitive question: What $\mu$ would assign the highest probability to the data $\{y^{(i)}\}_{i=1}^N$?

# Maximum likelihood

- Intuitive question: What $\mu$ would assign the highest probability to the data $\{y^{(i)}\}_{i=1}^{N}$?

$$\mu* = \mu^{\text{MLE}} = \arg\max_{\mu} P(\{y^{(i)}\}_{i=1}^{N} \,|\, \mu, \sigma^2) = \arg\max_{\mu} \prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(y^{(i)} - \mu)^2\right)$$

- This criterion to select model parameters based on highest probability is referred to as **maximum likelihood estimation** (MLE)

- MLE is a cornerstone of much of statistics and machine learning

- Find MLE of $\mu$ the same way as in our deterministic approach: **optimize**
  - Differentiate
  - Set to zero
  - Solve for $\mu$

# MLE regression with Gaussian noise

- Let's revisit our regression problem
- Inputs and targets are related via the equation

$$\boldsymbol{y}^{(i)} = \boldsymbol{\theta}^T \boldsymbol{x}^{(i)} + \boldsymbol{\epsilon}^{(i)}, \qquad i = 1,\ldots,N$$

  - Error term $\boldsymbol{\epsilon}^{(i)}$ captures unmodeled effects and noise
  - Error terms are independent and identically distributed (iid)

$$\boldsymbol{\epsilon}^{(i)} \sim \mathcal{N}(0,\sigma^2)$$

- Just as we can have different loss functions, we model $\boldsymbol{y}^{(i)}$ with different distributions
- Gaussian noise implies

$$p(\boldsymbol{y}^{(i)} \,|\, \boldsymbol{x}^{(i)}; \boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left( -\frac{(\boldsymbol{y}^{(i)} - \boldsymbol{\theta}^T \boldsymbol{x}^{(i)})^2}{2\sigma^2} \right)$$

# Probabilistic interpretation

- Very similar as in first example with MLE except that now want $\boldsymbol{\theta}$
- We now view $p(\boldsymbol{y} \,|\, \boldsymbol{x}; \boldsymbol{\theta})$ as a function of $\boldsymbol{\theta}$ (likelihood)

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{N} p(\boldsymbol{y}^{(i)} \,|\, \boldsymbol{x}^{(i)}; \boldsymbol{\theta})$$

where we used the independence assumption on $\boldsymbol{\epsilon}^{(i)}$

- **Principle of maximum likelihood** tells us to choose $\boldsymbol{\theta}$ that maximizes probability of data, i.e., that maximizes $L$

$$\max_{\boldsymbol{\theta}} L(\boldsymbol{\theta})$$

# Maximum likelihood estimation

board

# Maximum likelihood estimation

Instead of $L$, maximize the log-likelihood $\log L$

$$
\begin{aligned}
\ell(\theta) &= \log L(\theta) \\
&= \log \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left( -\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2} \right) \\
&= \sum_{i=1}^{N} \log \frac{1}{\sqrt{2\pi}\sigma} \exp\left( -\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2} \right) \\
&= N \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{\sigma^2}\frac{1}{2} \sum_{i=1}^{N} (y^{(i)} - \theta^T x^{(i)})^2
\end{aligned}
$$

Gives the very same optimum $\boldsymbol{\theta}^*$ as minimizing least-squares costs

# Predictions

- After we found $\boldsymbol{\theta}^*$ with MLE, we can make a prediction for new point $\boldsymbol{x}$

$$y = \boldsymbol{x}^T \boldsymbol{\theta}^*$$

- However: Probabilistic point of view even gives us a whole distribution

$$y \sim \mathcal{N}(\boldsymbol{x}^T \boldsymbol{\theta}^*, \sigma^2)$$

- The distribution accounts for noise and sometimes is more informative (with sufficient domain knowledge)

- Requires additionally fitting $\sigma^2$ $->$ with MLE

# Summary

- Point of view 1: Construct a loss function, then minimize empirical loss
- Point of view 2: Formulate a probabilistic model, then maximize likelihood
- Can we do the same for classification, rather than regression?

# Probabilistic approach to ***classification***

- Data
  - Inputs ("features") $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(N)} \in \mathscr{X} \subset \mathbb{R}^n$
  - Outputs ("targets") $\boldsymbol{y}^{(1)}, \ldots, \boldsymbol{y}^{(N)} \in \mathscr{Y} = \{0,1\}$

- Training data set $\mathscr{D} = \{(\boldsymbol{x}^{(1)}, \boldsymbol{y}^{(1)}), \ldots, (\boldsymbol{x}^{(N)}, \boldsymbol{y}^{(N)})\}$

- Gaussian noise model doesn't make much sense for classification because only have 0 and 1, rather than real values

- Typical distribution for binary data: Bernoulli (coin flip)

# Fit with principle of maximum likelihood

- Use principle of maximum likelihood to find $\boldsymbol{\theta}^*$
- Probabilistic assumption: Model $h_{\boldsymbol{\theta}}(x)$ gives the probability that y is 1, i.e.,

$$p(y = 1 \,|\, x; \boldsymbol{\theta}) = h_{\boldsymbol{\theta}}(x)$$

This also means (remember $h_{\boldsymbol{\theta}}(x) \in [0,1]$)

$$p(y = 0 \,|\, x; \boldsymbol{\theta}) = 1 - h_{\boldsymbol{\theta}}(x)$$

- Write more compactly as (because $y \in \{0,1\}$)

$$p(y \,|\, x; \boldsymbol{\theta}) = (h_{\boldsymbol{\theta}}(x))^y (1 - h_{\boldsymbol{\theta}}(x))^{(1-y)}$$

- (Remember that we modeled $y$ with a Gaussian distribution in the earlier regression problem. We now model $y$ with Bernoulli)

# Logistic regression

- Cannot simply use $h_\theta(x) = x^T\theta$ because need values between 0-1
- Therefore, transform $x^T\theta$ into $[0,1]$ with logistic function

$$h_\theta(x) = g(\theta^T x) = \frac{1}{1 + \mathrm{e}^{-\theta^T x}}$$

where

$$g(z) = \frac{1}{1 + \mathrm{e}^{-z}}$$

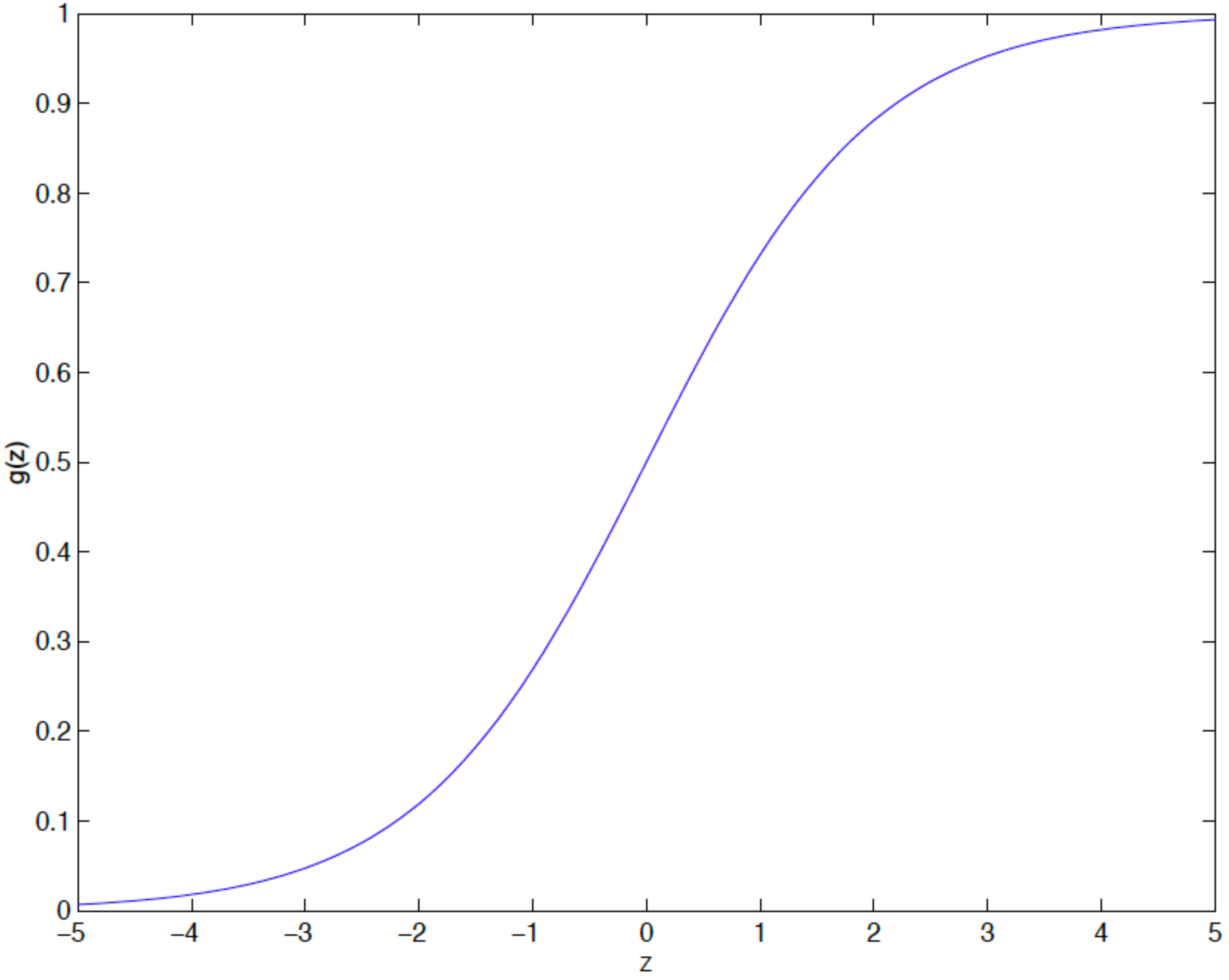
is called the logistic function or the sigmoid function

# Sigmoid

- $g(z)$ tends to 0 for $z \rightarrow -\infty$

- $g(z)$ tends to 1 for $z \rightarrow \infty$

- Ensures that $h_{\boldsymbol{\theta}}(x) \in [0,1]$




[Andrew Ng]

# Derivation of gradient descent update for logistic regression

board