CS189: Intro to Machine Learning
Summer 2018

Lecture 3: Review of probability. Probabilistic regression.

Josh Tobin
UC Berkeley EECS

Announcements

- Slides and ipynb's posted
- Location of tomorrow's review lecture may change. Keep an eye on Piazza!
- HW0 due tonight at 10pm on gradescope.

Some motivation

Ilya Sutskever

Head of research at OpenAI

One of the inventors of AlexNet

One of the inventors of seq-to-seq models

One of the inventors of dropout

…etc

Some motivation

“If I were talking to someone in an intro to machine learning class right now, my advice to them would be to hold on for dear life and enjoy the ride”

Outline for today

- Proof of ridge regression soln
- Probability
- Regression from a probabilistic perspective

Optimization motivation

	OLS	Ridge
Optimization problem	$\min_w \|\|Xw - y\|\|_2^2$	$\min_w \|\|Xw - y\|\|_2^2 + \lambda \|\|w\|\|_2^2$
Solution	$\hat{w} = (X^T X)^{-1} X^T y$	$\hat{w} = (X^T X + \lambda I)^{-1} X^T y$

Optimization motivation

Hyperparameter

$$\min_w ||Xw - y||_2^2 + \lambda ||w||_2^2$$

Training error

Keep weights small

$$\hat{w} = (X^T X + \lambda I)^{-1} X^T y$$

Outline for today

- Proof of ridge regression soln
- **Probability**
- Regression from a probabilistic perspective

Why probability theory for ML?

- Lots of events we want to model are uncertain
  - Knowable, but uncertain: “Will it rain today?”
  - Unknowable, uncertain: “Is my friend happy?”
- Even if the events are certain, observations may be noisy, e.g., sensor data
- Probability theory gives us a formal language to reason about uncertainty

Axioms for probability theory

1. $0 \leq p(A) \leq 1$

2. $p(\Omega) = 1 \quad p(\phi) = 0$

3. $p(A \cup B) = p(A) + p(B) - p(A \cap B)$

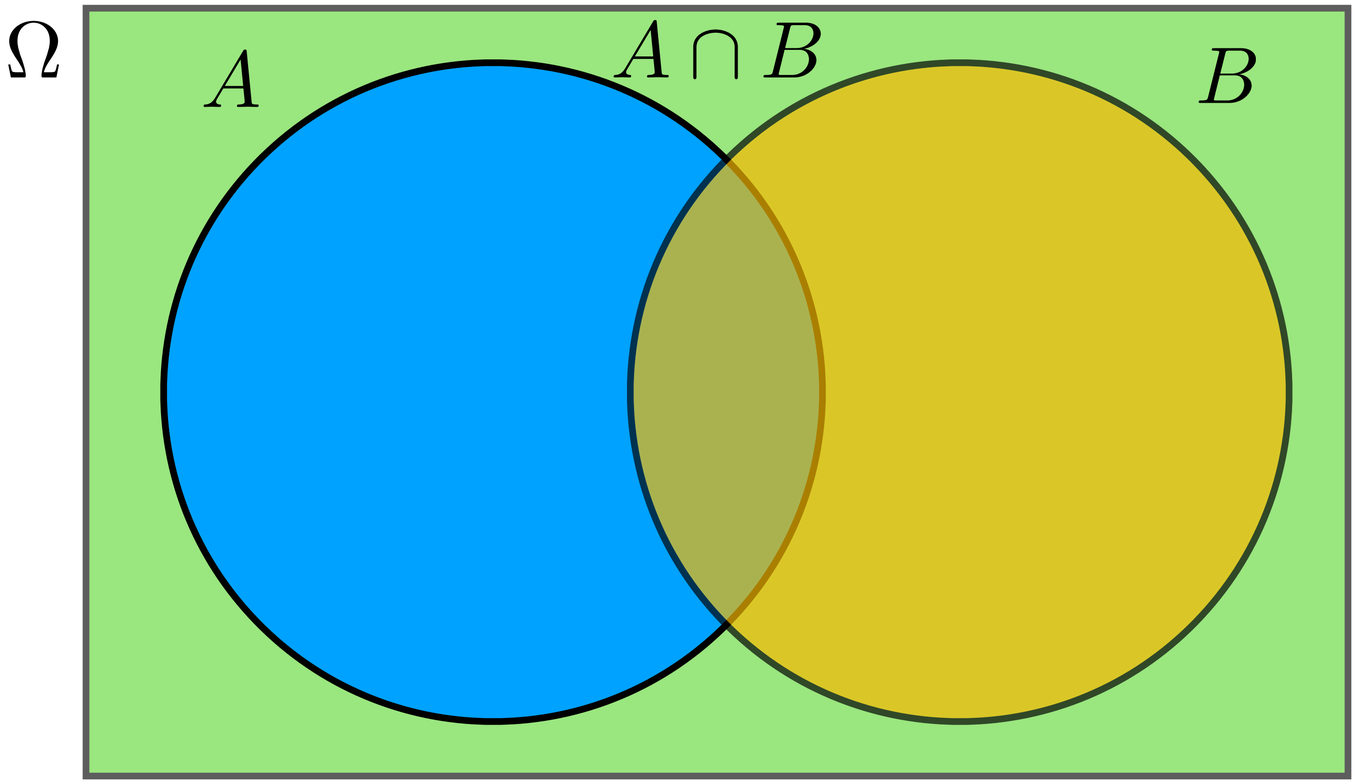
$p(A)$ Probability that the outcome w is an element of the set of possible outcomes A. A is often called an *event*

Ω set of all possible outcomes

ϕ empty set

A closer look at axiom 3

$$p(A \cup B) = p(A) + p(B) - p(A \cap B)$$



Discrete random variables

A **random variable** X is a variable that can take on a finite number of values $\{x_1, \cdots, x_n\}$

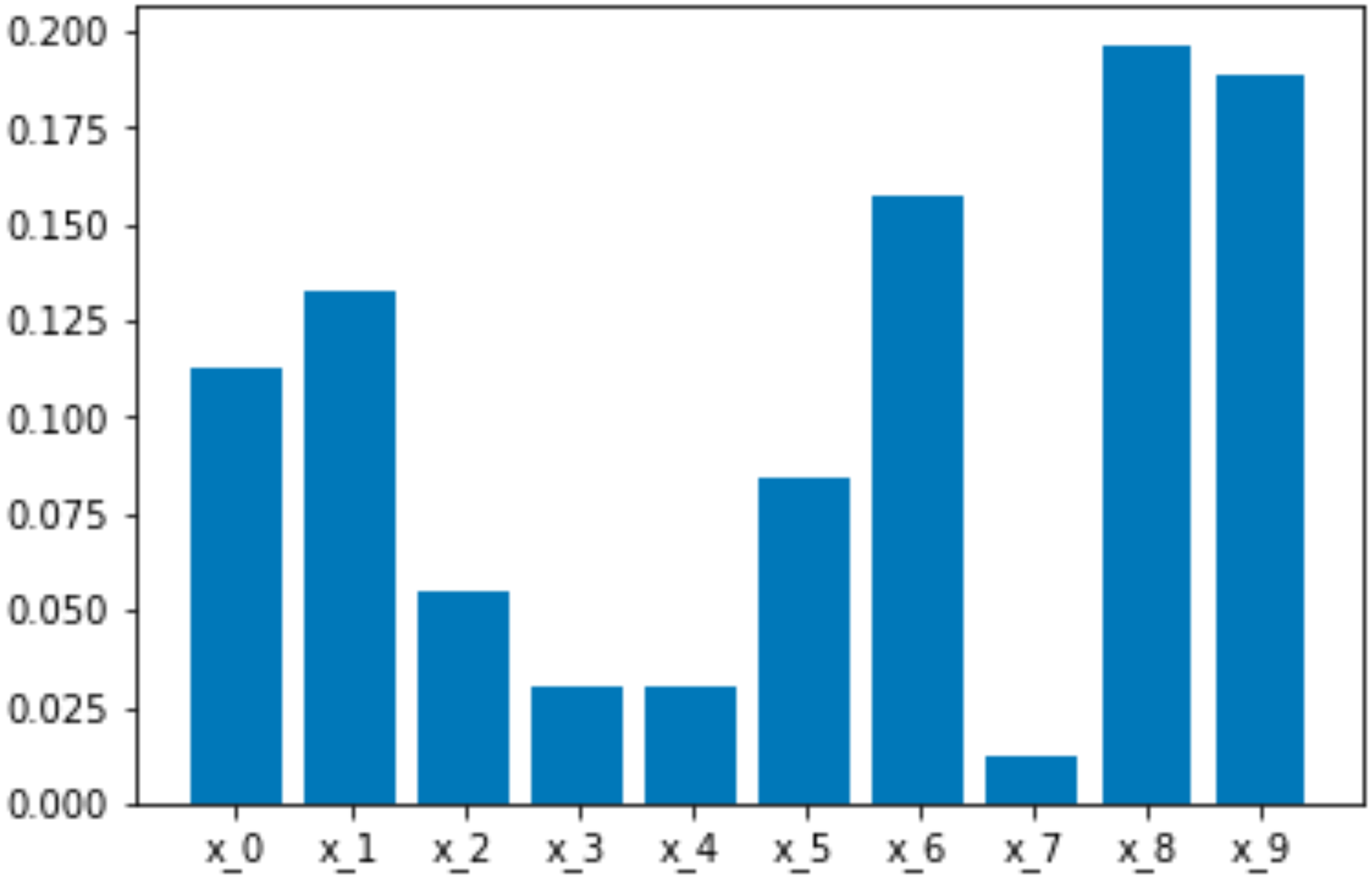
$p(X = x_i)$ or $p(x_i)$ is the probability that the random variable X takes on value xi

$p(\cdot)$ is called the **probability mass function**

e.g., X models the outcome of a coin flip.
x1 = heads, x2 = tails

$$p(x_1) = 0.5, \; p(x_2) = 0.5$$

Probability mass functions for discrete RVs



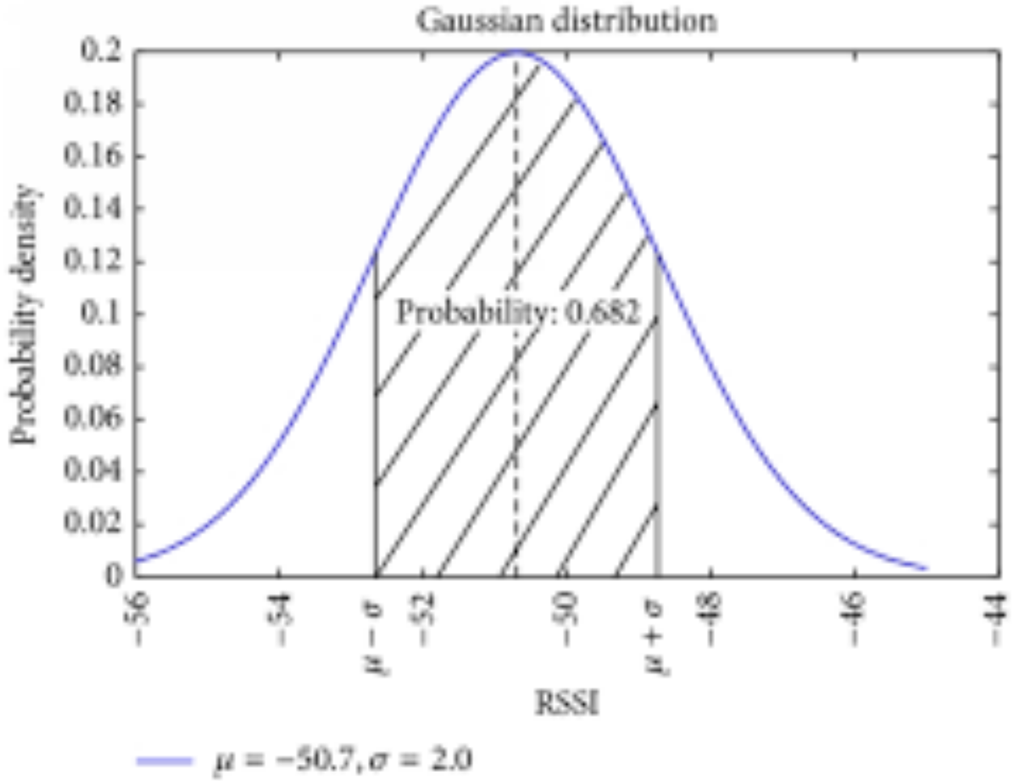
Continuous random variables

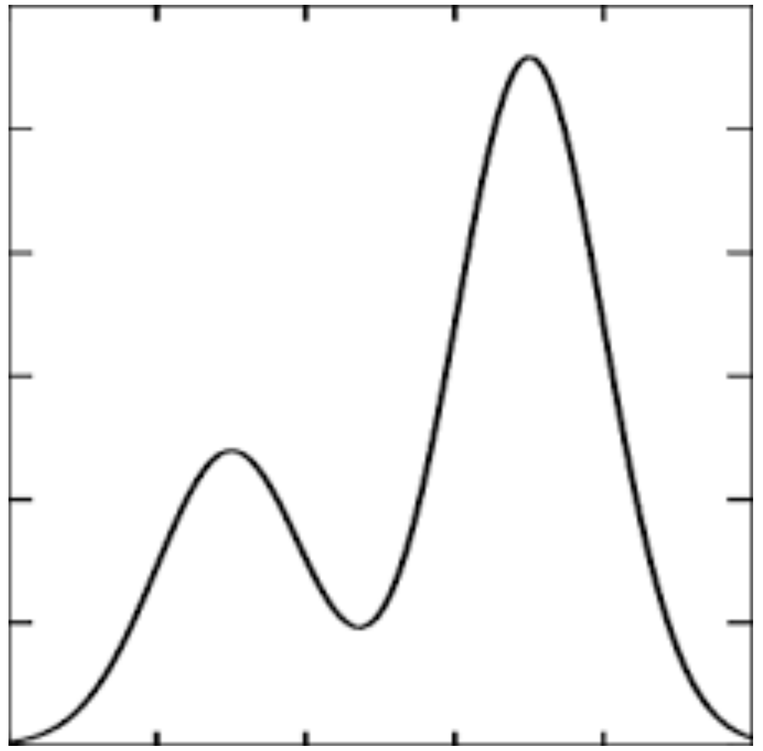
A **continuous random variable** X takes on values in a continuum

$p(X = x)$ or $p(x)$ is a **probability density function**

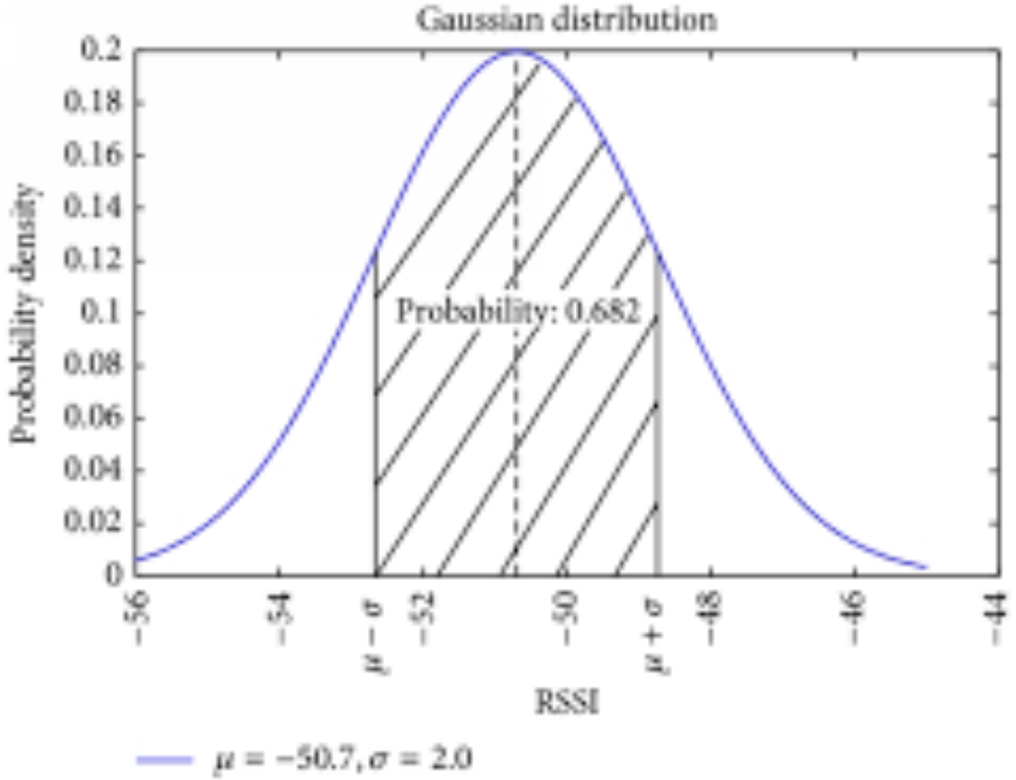
$$p\left(x \in (a, b)\right) = \int_a^b p(x)dx$$

Probability density functions





Normal distribution



"Standard" normal distribution

$$p(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

Parameterized normal distribution

$$p(x; \theta, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)}$$

Joint probability

$$p(X = x \text{ and } Y = y) = p(x, y)$$

If x and y are independent, $p(x, y) = p(x)p(y)$

Conditional probability

$p(x \mid y)$ is the probability of x *given* y

If x and y are independent, $p(x \mid y) = p(x)$

Conditional and joint probability are related:

$$p(x \mid y) = p(x, y)/p(y)$$

$$p(x, y) = p(x \mid y)p(y)$$

Law of total probability

Discrete case

$$\sum_x p(x) = 1$$

$$p(x) = \sum_y p(x, y)$$

$$p(x) = \sum_y p(x \mid y)p(y)$$

Continuous case

$$\int p(x)dx = 1$$

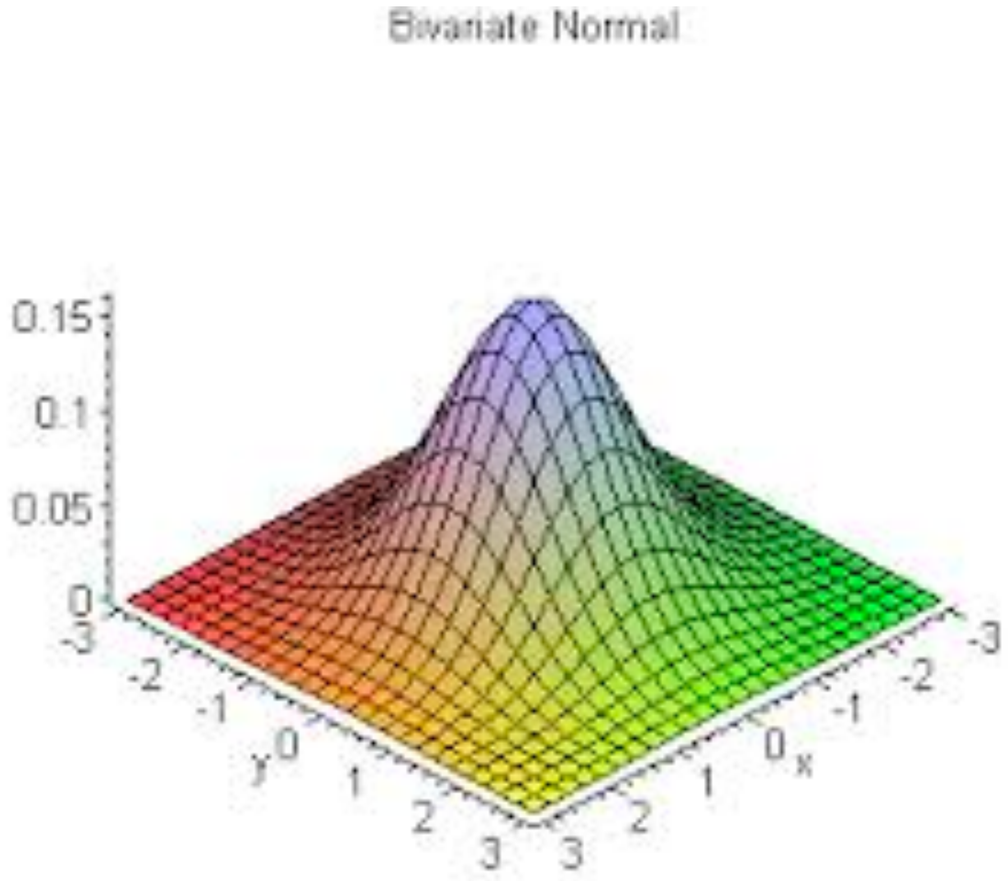
$$p(x) = \int p(x, y)dy$$

$$p(x) = \int p(x \mid y)p(y)dy$$

Probability tables

		X			
		0	1	2	3	$p(Y=y)$
Y	0	0.15	0.1	0.0875	0.0375	0.375
	1	0.1	0.175	0.1125	0	0.3875
	2	0.0875	0.1125	0	0	0.2
	3	0.0375	0	0	0	0.0375
	$p(X=x)$	0.375	0.3875	0.2	0.0375	1

Visualizing joint PDFs



Bayes rule

Variable to estimate

$$p(x \mid y) = \frac{p(y \mid x)p(x)}{p(y)}$$

Bayes rule

Variable to measure

$$p(x \mid y) = \frac{p(y \mid x)p(x)}{p(y)}$$

Bayes rule

Prior estimate of x

$$p(x \mid y) = \frac{p(y \mid x)p(x)}{p(y)}$$

Bayes rule

Model relating x & y

$$p(x \mid y) = \frac{p(y \mid x)p(x)}{p(y)}$$

Bayes rule

$$p(x \mid y) = \frac{p(y \mid x)p(x)}{p(y)}$$

Normalizing constant

$$\sum_x p(y \mid x)$$

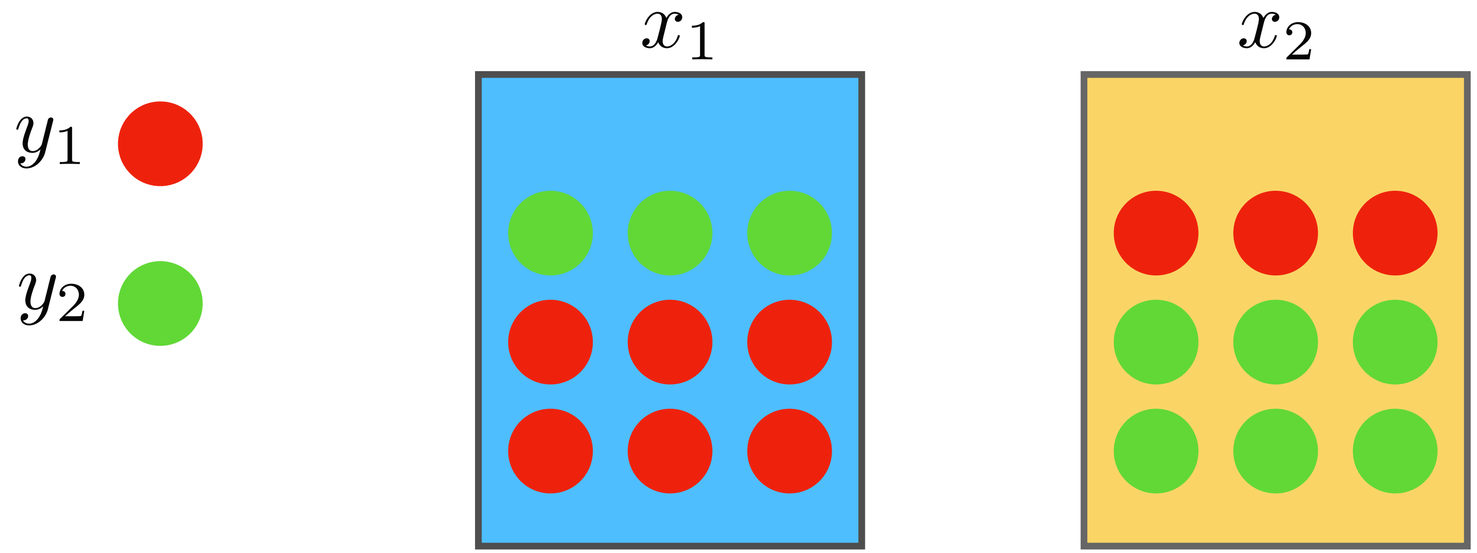
Bayes rule

$$p(x \mid y) = \frac{p(y \mid x)p(x)}{p(y)}$$

$$\text{posterior} = \frac{\text{likelihood} \cdot \text{prior}}{\text{evidence}}$$

$$p(x, y) = p(x \mid y)p(y) = p(y \mid x)p(x)$$

Bayes rule example



A person **randomly picks a ball** from one of two boxes. They are **equally likely to pick box 1 and box 2**. If a red ball was drawn, what's the **probability they picked box 1?**

$$p(X = x_1 \mid Y = y_1) = \frac{p(Y = y_1 \mid X = x_1)p(X = x_1)}{p(Y = y_1)} = \frac{2/3 \cdot 0.5}{2/3 * 0.5 + 1/3 * 0.5} = \frac{2}{3}$$

Outline for today

- Proof of ridge regression soln
- Probability
- **Regression from a probabilistic perspective**

Four levels for ML problems

1. Data & application
2. Model
3. Optimization problem
4. Optimization algorithm

Recall two optimization problems

OLS $\min_w ||Xw - y||_2^2$

Ridge regression $\min_w ||Xw - y||_2^2 + \lambda ||w||_2^2$

Why are these the right optimization problems?

Where we're going

1. Define a probabilistic model for our data
2. Define what it means for a model to be "good" probabilistically
3. Do some math to show that the "best" linear model probabilistically corresponds to the OLS estimate

A probabilistic model for supervised learning

Data $X, y = \{(\vec{x}_i, y_i)\}_{i=1}^{n}$

Underlying model $f : \vec{x} \rightarrow f(\vec{x})$

What is z?

How are they related? $y_i = f(\vec{x}_i) + z_i$

A probabilistic model for supervised learning

$$y_i = f(\vec{x}_i) + z_i$$

What is z?

- Mean = 0
- All z_i are *independent*
- Each z_i comes from the same distribution (*identically distributed*)

i.i.d assumptions

(independent identically distributed)

A probabilistic model for supervised learning

$$y_i = f(\vec{x}_i) + z_i$$

What is z?

- zero-mean, i.i.d
- Assume a Gaussian distribution

$$z_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$$

Therefore:

$$y_i \overset{\text{iid}}{\sim} \mathcal{N}(f(\vec{x}_i), \sigma^2)$$

Where we're going

1. Define a probabilistic model for our data
2. Define what it means for a model to be "good" probabilistically
3. Do some math to show that the "best" linear model probabilistically corresponds to the OLS estimate

Maximum Likelihood Estimation

f_θ = *Family* of models, e.g., $f_\theta(x) = \theta^T x$

Each theta gives us an implied value for the probability (i.e., *likelihood*) of the data we observed

$$\mathcal{L}(\theta; \mathcal{D}) = p(\text{data} = \mathcal{D} \mid \text{model} = f_\theta)$$

The "*best*" model is the one that maximizes the likelihood

$$\hat{\theta}_{\text{MLE}} = \text{argmax}_\theta \mathcal{L}(\theta; \mathcal{D})$$

Analyzing the MLE solution

The "*best*" model is the one that maximizes the likelihood

$$\hat{\theta}_{\text{MLE}} = \text{argmax}_{\theta} \mathcal{L}(\theta; \mathcal{D})$$

$$\hat{\theta}_{\text{MLE}} = \text{argmax}_{\theta} \mathcal{L}(\theta; X, y) = p(y_1, \cdots, y_n \mid x_1, \cdots, x_n, \theta, \sigma^2)$$

$$\implies \hat{\theta}_{\text{MLE}} = \text{argmax}_{\theta} \prod_{i=1}^{n} p(y_i \mid \vec{x}_i, \theta, \sigma^2)$$

Insight: argmax p <-> argmax logp

$$\hat{\theta}_{\text{MLE}} = \text{argmax}_\theta \mathcal{L}(\theta; X, y) = \text{argmax}_\theta \log \mathcal{L}(\theta; X, y) = \text{argmax}_\theta l(\theta; X, y)$$

$$\hat{\theta}_{\text{MLE}} = \text{argmax}_\theta \prod_{i=1}^{n} p(y_i \mid \vec{x}_i, \theta, \sigma^2)$$

$$\implies \hat{\theta}_{\text{MLE}} = \text{argmax}_\theta \log \prod_{i=1}^{n} p(y_i \mid \vec{x}_i, \theta, \sigma^2)$$

$$\implies \hat{\theta}_{\text{MLE}} = \text{argmax}_\theta \sum_{I=1}^{n} \log p(y_i \mid \vec{x}_i, \theta, \sigma^2)$$

MLE justifies OLS

$$\hat{\theta}_{\text{MLE}} = \text{argmin}_\theta \sum_{i=1}^{n} (y_i - f_\theta(x_i))^2$$

Linear regression model:

$$\hat{\theta}_{\text{MLE}} = \text{argmin}_\theta \sum_{i=1}^{n} (y_i - \theta^T x_i)^2 = ||y - X\theta||_2^2$$

Next time

- Maximum a posteriori estimates
  (i.e., what's the justification for the ridge regression optimization problem?)
- Bias / variance tradeoff