STATISTICS

EXAMPLES

Bernoulli(p)

$$f(0) = 1 - p, f(1) = p$$
$$\mu = p. \quad \sigma^2 = p(1-p) = pq$$
$$m(t) = pe^t + q$$

Poisson(λt)

$$f(x) = \frac{1}{x!}(\lambda t)^x e^{-\lambda t}, \text{ for } x = 0, 1, \ldots$$
$$\mu = \lambda t. \quad \sigma^2 = \lambda t$$
$$m(s) = e^{\lambda t(e^s - 1)}$$

Binomial(n, p)

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}, \text{ for } x = 0, 1, \ldots, n$$
$$\mu = np. \quad \sigma^2 = np(1-p) = npq$$
$$m(t) = (pe^t + q)^n$$

Geometric(p)

$$f(x) = q^{x-1}p, \text{ for } x = 1, 2, \ldots$$
$$\mu = \frac{1}{p}. \quad \sigma^2 = \frac{1-p}{p^2}$$
$$m(t) = \frac{pe^t}{1 - qe^t}$$

EXAMPLES

$$\text{UNIFORM}(a,b)$$
$$f(x) = \frac{1}{b-a}, \text{ for } x \in [a,b]$$
$$\mu = \frac{a+b}{2}. \quad \sigma^2 = \frac{(b-a)^2}{12}$$
$$m(t) = \frac{e^{bt} - e^{at}}{t(b-a)}$$

$$\text{NORMAL}(\mu, \sigma^2)$$
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \text{ for } x \in \mathbf{R}$$
$$\mu = \mu. \quad \sigma^2 = \sigma^2$$
$$m(t) = \exp(\mu t + t^2\sigma^2/2)$$

$$\text{EXPONENTIAL}(\lambda)$$
$$f(x) = \lambda e^{-\lambda x}, \text{ for } x \in [0,\infty)$$
$$\mu = 1/\lambda. \quad \sigma^2 = 1/\lambda^2$$
$$m(t) = (1 - t/\lambda)^{-1}$$

$$\text{CHISQUARED}(\nu)$$
$$f(x) = \frac{x^{\nu/2-1}e^{x/2}}{2^{\nu/2}\Gamma(\nu/2)}, \text{ for } x \geq 0$$
$$\mu = \nu. \quad \sigma^2 = 2\nu$$
$$m(t) = (1-2t)^{\nu/2}$$

EXPECTATION

The expectation (mean) of a random variable X can be expressed as

$$E(X) = X_{\text{mean}} = \sum_{x \in \mathcal{X}} xP(X = x). \tag{4}$$

The variance and covariance can be defined therefore in terms of the expectation, where

Standard deviation is the square root of the variance.

$$Var(X) = E((X - E(X))^2) = E(X^2) - E(X)^2, \tag{5}$$

$$Cov(X, Y) = E(XY) - E(X)E(Y). \tag{6}$$

Conditional expectations and variances follow from the conditional distributions

$$E(X \mid Y = y) = \sum_{x \in \mathcal{X}} xP(X = x \mid Y = y). \tag{7}$$

EXPECTATION AND VARIANCE OF IMPORTANT RANDOM VARIABLES

Distribution	Mean	Variance
Point mass at a	a	0
Bernoulli(p)	p	$p(1-p)$
Binomial(n,p)	np	$np(1-p)$
Geometric(p)	$1/p$	$(1-p)/p^2$
Poisson(λ)	λ	λ
Uniform(a,b)	$(a+b)/2$	$(b-a)^2/12$
Normal(μ,σ^2)	μ	σ^2
Exponential(β)	β	β^2
Gamma(α,β)	$\alpha\beta$	$\alpha\beta^2$
Beta(α,β)	$\alpha/(\alpha+\beta)$	$\alpha\beta/((\alpha+\beta)^2(\alpha+\beta+1))$
t_ν	0 (if $\nu > 1$)	$\nu/(\nu-2)$ (if $\nu > 2$)
χ^2_p	p	$2p$
Multinomial(n,p)	np	see below
Multivariate Normal(μ,Σ)	μ	Σ

$n(I-pp^T)$

PROBABILITY TABLES

You can use a probability table to represent the probability information of a 2-variable discrete distribution clearly.

	Coin 2: Heads	Coin 2: Tails
Coin 1: Heads	1/8	1/2
Coin 1: Tails	1/4	1/8

INDEPENDENCE

Two random variables are independent when the probability distribution of one random variable does not affect the other. More concretely, two random variables X and Y are independent, that is, $X \perp Y$, if and only if

$$P(X = x, Y = y) = P(X = x)P(Y = y), \quad \text{for all } x \in \mathcal{X}, y \in \mathcal{Y}. \tag{8}$$

If X and Y are continuous with joint density function $f_{X,Y}(x, y)$, then the above condition reduces to finding functions $h(x)$ and $g(y)$ such that

$$f_{X,Y}(x, y) = h(x)g(y). \tag{9}$$

CONDITIONAL INDEPENDENCE

Two random variables X and Y are conditionally independent given a third variable Z, denoted as $X \perp Y \mid Z$, if and only if

$$P(X = x, Y = y \mid Z = z) = P(X = x \mid Z = z)P(Y = y \mid Z = z), \quad (10)$$

for all $x \in \mathcal{X}, y \in \mathcal{Y}, z \in \mathcal{Z}$.
This is equivalent to saying

$$P(X = x \mid Y = y, Z = z) = P(X = x \mid Z = z).$$

Note that $X \perp Y \mid Z$ does not imply that $X \perp Y$, and vice versa.

C.I. RELATIONS

- Symmetry:

$$X \perp Y \,|\, Z \implies Y \perp X \,|\, Z$$

- Decomposition:

$$X \perp Y, W \,|\, Z \implies X \perp Y \,|\, Z \quad (\text{and } X \perp W \,|\, Z)$$

- Weak union:

$$X \perp Y, W \,|\, Z \implies X \perp Y \,|\, Z, W \quad (\text{and } X \perp W \,|\, Y, Z)$$

- Contraction:

$$X \perp Y \,|\, Z \text{ and } X \perp W \,|\, Y, Z \implies X \perp Y, W \,|\, Z$$

MAXIMUM LIKELIHOOD

- The Maximum Likelihood Estimation (MLE) is a method of estimating the parameters of a model. This estimation method is one of the most widely used.
- The method of maximum likelihood selects the set of values of the model parameters that maximizes the likelihood function. Intuitively, this maximizes the "agreement" of the selected model with the observed data.
- The Maximum-likelihood Estimation gives an unified approach to estimation.

MAXIMUM LIKELIHOOD

Definition

This joint probability is a function of θ (the unknown parameter) and corresponds to the **likelihood of the sample** $\{x_1, .., x_N\}$ denoted by

$$L_N(\theta; x_1.., x_N) = \Pr((X_1 = x_1) \cap ... \cap (X_N = x_N))$$

You can break up the likelihood into products because samples are usually independent and identically distributed (i.i.d.).

Question: What value of θ would make this **sample most probable**?

EXAMPLE: BERNOULLI

Example 1. A coin is flipped 100 times. Given that there were 55 heads, find the maximum likelihood estimate for the probability p of heads on a single toss.

Before actually solving the problem, let's establish some notation and terms.

We can think of counting the number of heads in 100 tosses as an experiment. For a given value of p, the probability of getting 55 heads in this experiment is the binomial probability

$$P(55 \text{ heads}) = \binom{100}{55} p^{55}(1-p)^{45}.$$

The probability of getting 55 heads depends on the value of p, so let's include p in by using the notation of conditional probability:

$$P(55 \text{ heads} \,|\, p) = \binom{100}{55} p^{55}(1-p)^{45}.$$

You should read $P(55 \text{ heads} \,|\, p)$ as:

'the probability of 55 heads given p,'

or more precisely as

'the probability of 55 heads given that the probability of heads on a single toss is p.'

EXAMPLE: BERNOULLI

- Likelihood, or likelihood function: this is $P(\text{data}\,|\,p)$. Note it is a function of both the data and the parameter p. In this case the likelihood is

$$P(55 \text{ heads}\,|\,p) = \binom{100}{55} p^{55}(1-p)^{45}.$$

Definition: Given data the maximum likelihood estimate (MLE) for the parameter p is the value of p that maximizes the likelihood $P(\text{data}\,|\,p)$. That is, the MLE is the value of p for which the data is most likely.

EXAMPLE: BERNOULLI

answer: For the problem at hand, we saw above that the likelihood

$$P(55 \text{ heads} \,|\, p) = \binom{100}{55} p^{55}(1-p)^{45}.$$

We'll use the notation $\hat{p}$ for the MLE. We use calculus to find it by taking the derivative of the likelihood function and setting it to 0.

$$\frac{d}{dp} P(\text{data} \,|p) = \binom{100}{55} (55p^{54}(1-p)^{45} - 45p^{55}(1-p)^{44}) = 0.$$

Solving this for p we get

$$55p^{54}(1-p)^{45} = 45p^{55}(1-p)^{44}$$
$$55(1-p) = 45p$$
$$55 = 100p$$
$$\text{the MLE is } \hat{p} = .55$$

This is known as the estimator.

EXAMPLE: POISSON

Example

Suppose that $X_1, X_2, \cdots, X_N$ *are i.i.d. discrete random variables, such that* $X_i \sim Pois(\theta)$ *with a* **pmf** (probability mass function) defined as:

$$\Pr(X_i = x_i) = \frac{\exp(-\theta)\,\theta^{x_i}}{x_i!}$$

where θ is an unknown parameter to estimate.

EXAMPLE: POISSON

Question: What is the probability of observing the **particular sample** $\{x_1, x_2, .., x_N\}$, assuming that a Poisson distribution with as yet unknown parameter θ generated the data?

This probability is equal to

$$\Pr\left((X_1 = x_1) \cap ... \cap (X_N = x_N)\right)$$

EXAMPLE: POISSON

Since the variables X_i are *i.i.d.* this joint probability is equal to the product of the marginal probabilities

$$\Pr\left((X_1 = x_1) \cap \ldots \cap (X_N = x_N)\right) = \prod_{i=1}^{N} \Pr\left(X_i = x_i\right)$$

Given the pmf of the Poisson distribution, we have:

$$\begin{aligned} \Pr\left((X_1 = x_1) \cap \ldots \cap (X_N = x_N)\right) &= \prod_{i=1}^{N} \frac{\exp\left(-\theta\right)\theta^{x_i}}{x_i!} \\ &= \exp\left(-\theta N\right) \frac{\theta^{\sum_{i=1}^{N} x_i}}{\prod_{i=1}^{N} x_i!} \end{aligned}$$

EXAMPLE: POISSON

Consider maximizing the likelihood function $L_N(\theta; x_1.., x_N)$ with respect to θ. Since the log function is monotonically increasing, we usually maximize $\ln L_N(\theta; x_1.., x_N)$ instead. In this case:

$$\ln L_N(\theta; x_1.., x_N) = -\theta N + \ln(\theta) \sum_{i=1}^{N} x_i - \ln\left(\prod_{i=1}^{N} x_i!\right)$$

$$\frac{\partial \ln L_N(\theta; x_1.., x_N)}{\partial \theta} = -N + \frac{1}{\theta} \sum_{i=1}^{N} x_i$$

$$\frac{\partial^2 \ln L_N(\theta; x_1.., x_N)}{\partial \theta^2} = -\frac{1}{\theta^2} \sum_{i=1}^{N} x_i < 0$$

EXAMPLE: POISSON

Under suitable regularity conditions, the maximum likelihood estimate (estimator) is defined as:

$$\widehat{\theta} = \underset{\theta \in \mathbb{R}^+}{\arg\max} \ln L_N(\theta; x_1.., x_N)$$

$$FOC: \left. \frac{\partial \ln L_N(\theta; x_1.., x_N)}{\partial \theta} \right|_{\widehat{\theta}} = -N + \frac{1}{\widehat{\theta}} \sum_{i=1}^{N} x_i = 0$$

$$\Longleftrightarrow \widehat{\theta} = (1/N) \sum_{i=1}^{N} x_i$$

$$SOC: \left. \frac{\partial^2 \ln L_N(\theta; x_1.., x_N)}{\partial \theta^2} \right|_{\widehat{\theta}} = -\frac{1}{\widehat{\theta}^2} \sum_{i=1}^{N} x_i < 0$$

$\widehat{\theta}$ is a maximum.

CONTINUOUS MLE

Continuous variables

- The reference to the probability of observing the given sample is not exact in a continuous distribution, since a particular sample has probability zero. Nonetheless, the principle is the same.

- The likelihood function then corresponds to the pdf associated to the **joint distribution** of $(X_1, X_2, .., X_N)$ evaluated at the point $(x_1, x_2, .., x_N)$:

$$L_N(\theta; x_1.., x_N) = f_{X_1,..,X_N}(x_1, x_2, .., x_N; \theta)$$

CONTINUOUS MLE

Continuous variables

- If the random variables $\{X_1, X_2, .., X_N\}$ are $i.i.d.$ then we have:

$$L_N(\theta; x_1.., x_N) = \prod_{i=1}^{N} f_X(x_i; \theta)$$

where $f_X(x_i; \theta)$ denotes the pdf of the marginal distribution of X (or X_i since all the variables have the same distribution).

- The values of the parameters that maximize $L_N(\theta; x_1.., x_N)$ or its log are the maximum likelihood estimates, denoted $\widehat{\theta}(x)$.

REGRESSION

MACHINE LEARNING

Task

Performance

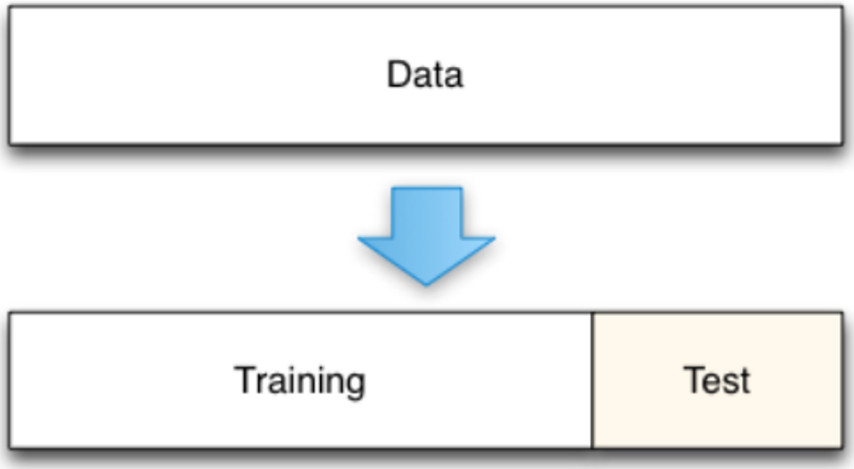
Experience

Algorithms that improve their performance
at some task with experience
– Tom Mitchell (1998)

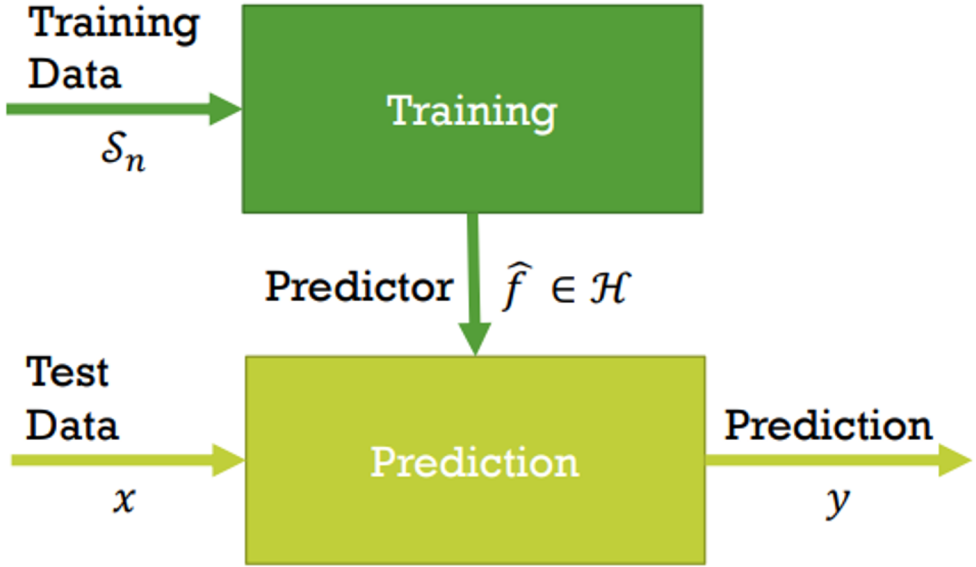
TRAINING DATA VS TEST DATA

Partition data into:

- Training data set $\mathcal{S}_n$
- Test data set $\mathcal{S}_*$



LEARNING AND PREDICTION



Assumption. Test data and training data are identically distributed.

TRAINING MODEL

$$\mathcal{S}_n = \left\{ \left(x^{(i)}, y^{(i)} \right) \mid i = 1, \ldots, n \right\}$$

- **Features/Inputs** $x^{(i)} = \left(x_1^{(i)}, \ldots, x_d^{(i)} \right)^\top \in \mathbb{R}^d$
- **Response/Output** $y^{(i)} \in \mathbb{R}$

TRAINING MODEL

Each f is a *predictor* or *hypothesis*

Model (or Hypothesis Class) $\mathcal{H}$

Set of *linear* functions $f: \mathbb{R}^d \rightarrow \mathbb{R}$

$$f(x; \theta, \theta_0) = \theta_d x_d + \cdots + \theta_1 x_1 + \theta_0 = \theta^\top x + \theta_0$$

Model Parameters

$$\theta \in \mathbb{R}^d, \theta_0 \in \mathbb{R}$$

TRAINING

Sometimes, we write $\mathcal{L}(\theta; S_n)$ instead of $\mathcal{L}(f; S_n)$

Training Loss/Objective

$$\mathcal{L}(f; S_n) = \frac{1}{n} \sum_{(x,y) \in S_n} \frac{1}{2} \big(y - f(x)\big)^2$$

Find predictor $\hat{f} \in \mathcal{H}$ that minimizes $\mathcal{L}(f; S_n)$.

Training Algorithm

Set gradient to zero, and solve equations.

Training is also sometimes called *Learning*.

GENERALIZATION

The goal of machine learning is to find a predictor $\hat{f} \in \mathcal{H}$ that generalizes well, i.e. that predicts well on test data $\mathcal{S}_*$.

TESTING

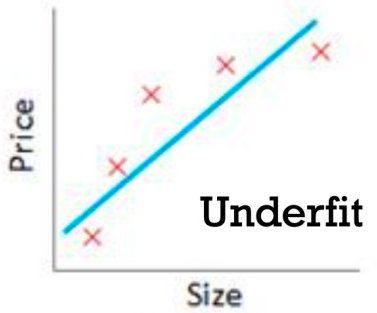
Sometimes, we write $\mathcal{R}(\hat{\theta}; \mathcal{S}_*)$ instead of $\mathcal{R}(\hat{f}; \mathcal{S}_*)$

Test Loss/Objective

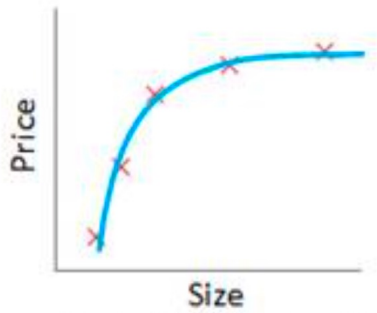
$$\mathcal{R}(\hat{f}; \mathcal{S}_*) = \frac{1}{n} \sum_{(x,y) \in \mathcal{S}_*} \frac{1}{2} \left(y - \hat{f}(x) \right)^2$$

We often use some test loss $\mathcal{R}(\hat{f}; \mathcal{S}_*)$ to measure how well a predictor $\hat{f}$ generalizes. The test loss can be different from the training loss $\mathcal{L}(f; \mathcal{S}_n)$.

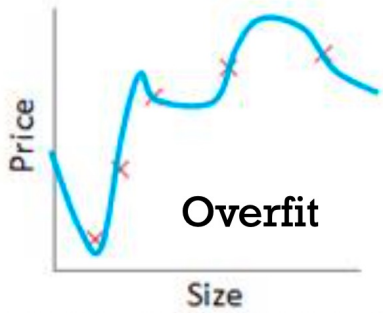
FITTING



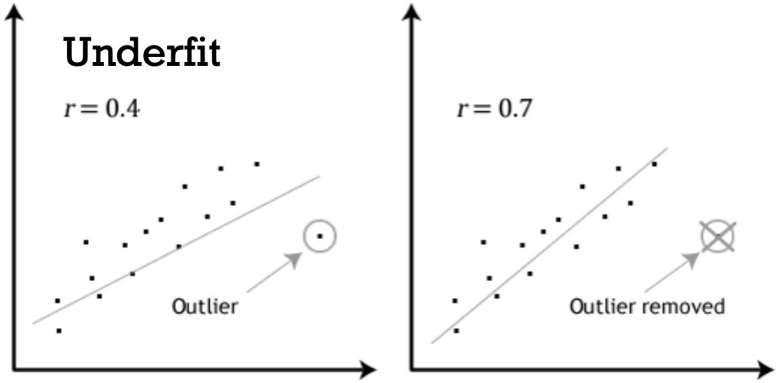
$\theta_0 + \theta_1 x$



$\theta_0 + \theta_1 x + \theta_2 x^2$



$\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$



FITTING

Overfitting. If model $\mathcal{H}$ is too big, then $\hat{f} \in \mathcal{H}$ performs

- well on training data, but poorly on test data.

Underfitting. If model $\mathcal{H}$ is too small, then $\hat{f} \in \mathcal{H}$ performs

- poorly on training data, and poorly on test data.

Finding a model with the right size is called model selection.

REGRESSION

Machine Learning
> Supervised Learning
> Regression

- **Task.** Find function $f: \mathbb{R}^d \rightarrow \mathbb{R}$ such that $y \approx f(x; \theta)$
- **Experience.** Training data $\left(x^{(1)}, y^{(1)}\right), \ldots, \left(x^{(n)}, y^{(n)}\right)$
- **Performance.** Prediction error $y - f(x; \theta)$ on test data

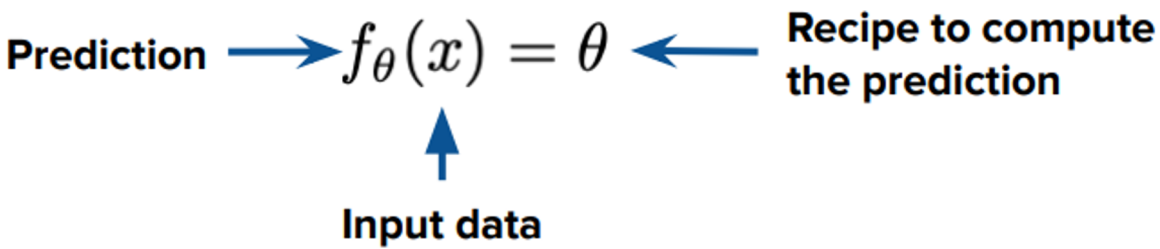
MODEL SELECTION

- Loss minimization framework useful for predictions too!
- Suppose we have a dataset of cars and we'd like to predict fuel efficiency (miles per gallon, or mpg):

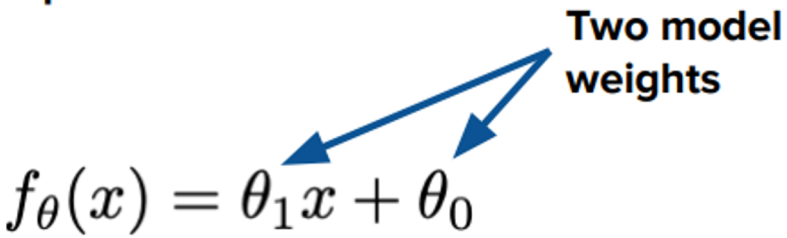
	mpg	cylinders	displacement	horsepower	...	acceleration	model_year	origin	name
0	18.0	8	307.0	130.0	...	12.0	70	usa	chevrolet chevelle malibu
1	15.0	8	350.0	165.0	...	11.5	70	usa	buick skylark 320
2	18.0	8	318.0	150.0	...	11.0	70	usa	plymouth satellite

MODEL SELECTION

- To make a prediction, we choose a **model**.
  - Takes input data and outputs a prediction.
- Constant model:



- Simple linear model:



MODEL LOSS

- Use x_i to denote what we use to make predictions
- Use y_i to denote what we're trying to predict
- But both x and y come from a single sample
- Idea: Pick the θ that minimizes the average loss between y in our sample and model predictions.

$$L(\theta, y_1, \ldots, y_n) = \frac{1}{n}\sum (y_i - f_\theta(x))^2$$

This is not the likelihood function!

CONSTANT MODEL

- Start simple: if constant model, how do we pick θ?

$$f_\theta(x) = \theta$$

- Intuition: pick θ to be close to most of the values in data

	mpg	cylinders	displacement	horsepower	...	acceleration	model_year	origin	name
0	18.0	8	307.0	130.0	...	12.0	70	usa	chevrolet chevelle malibu
1	15.0	8	350.0	165.0	...	11.5	70	usa	buick skylark 320
2	18.0	8	318.0	150.0	...	11.0	70	usa	plymouth satellite

CONSTANT MODEL LOSS

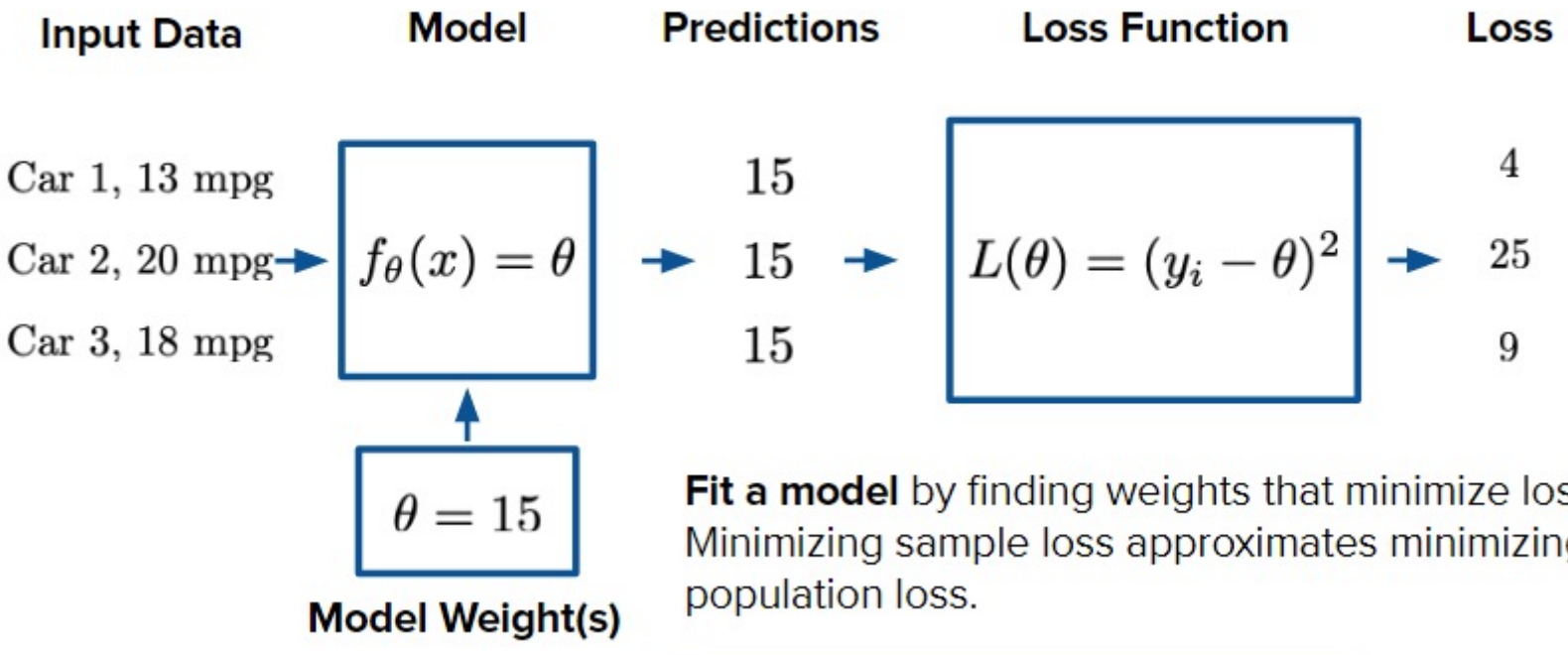
$$L(\theta, y_1, \ldots, y_n) = \frac{1}{n}\sum (y_i - f_\theta(x))^2$$

Since $f_\theta(x) = \theta$ for constant model:

$$L(\theta, y_1, \ldots, y_n) = \frac{1}{n}\sum (y_i - \theta)^2$$

- θ = sample mean is the best model parameter.
- So, for car MPGs, we set θ = mean(mpg)

THE MODELING PIPELINE



Fit a model by finding weights that minimize loss. Minimizing sample loss approximates minimizing population loss.

We choose what goes in the blue boxes!

LOSS FUNCTION EXAMPLES

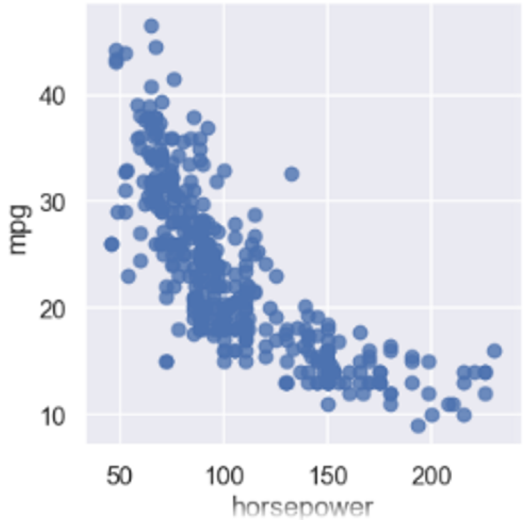
- Pick a model, pick a loss function, fit the model to sample.
- Preview of model and loss function combos:

Model	Loss Function	Technique Name
$\boldsymbol{\theta} \cdot \boldsymbol{x}$	$\frac{1}{n}\sum(y_i - f_\theta(\boldsymbol{x}))^2$	Least squares linear regression
$\boldsymbol{\theta} \cdot \boldsymbol{x}$	$\frac{1}{n}\sum(y_i - f_\theta(\boldsymbol{x}))^2 + \lambda\|\boldsymbol{\theta}\|^2$	Ridge regression
$\boldsymbol{\theta} \cdot \boldsymbol{x}$	$\frac{1}{n}\sum(y_i - f_\theta(\boldsymbol{x}))^2 + \lambda\|\|\boldsymbol{\theta}\|\|_{\ell_1}$	Lasso regression
$\boldsymbol{\theta} \cdot \boldsymbol{x}$	$\frac{1}{n}\sum\|y_i - f_\theta(\boldsymbol{x})\|$	Least absolute deviations
$\sigma(\boldsymbol{\theta} \cdot \boldsymbol{x})$	$\frac{1}{n}\sum[-y_i \ln f_\theta(\boldsymbol{x}) - (1-y_i)\ln(1-f_\theta(\boldsymbol{x}))]$	Logistic regression

LINEAR REGRESSION

- If we're trying to predict MPG, we can do better than a constant model by incorporating more information.
  - E.g. higher horsepowers have lower MPGs:

	mpg	cylinders	displacement	horsepower
0	18.0	8	307.0	130.0
1	15.0	8	350.0	165.0
2	18.0	8	318.0	150.0



SIMPLE LINEAR MODEL

- We want our predictions to depend on the input data x.
- Simple linear model:

$$f_\theta^*(x) = \theta_1^* x + \theta_0^* + \epsilon \qquad \epsilon \sim N(0, \sigma^2)$$

$$f_\theta(x) = \theta_1 x + \theta_0$$

- As usual, we can minimize the loss. This time, we have two parameters.

$$L(\theta_1, \theta_0, y_1, \ldots, y_n) = \frac{1}{n} \sum (y_i - f_\theta(x_i))^2$$

$$= \frac{1}{n} \sum (y_i - \theta_1 x_i - \theta_0)^2$$

SIMPLE LINEAR MODEL

$$L(\theta_1, \theta_0, y_1, \ldots, y_n) = \frac{1}{n}\sum(y_i - \theta_1 x_i - \theta_0)^2$$

$$\frac{\partial}{\partial\theta_1}L = \frac{1}{n}\sum 2(y_i - \theta_1 x_i - \theta_0)(-x_i)$$

$$\frac{\partial}{\partial\theta_0}L = \frac{1}{n}\sum 2(y_i - \theta_1 x_i - \theta_0)(-1)$$

- This ends up being a lot of algebra, so we'll skip to the answer.

SIMPLE LINEAR MODEL

Let r be the average of the products of x and y when both variables are measured in standard units.

$$\hat{\theta}_1 = r \cdot \frac{\text{SD of } y}{\text{SD of } x} \qquad \hat{\theta}_0 = \text{mean of } y - \hat{\theta}_1 \cdot \text{mean of } x$$

You can convert any sample set of a variable to standard units by subtracting the mean from each sample and then dividing by the standard deviation of the entire sample set of that variable.

MULTIVARIATE LINEAR REGRESSION

- Simple linear model uses one variable to predict:

$$f_\theta(x) = \theta_1 x + \theta_0$$

- Multivariable linear model uses ≥1 variable:

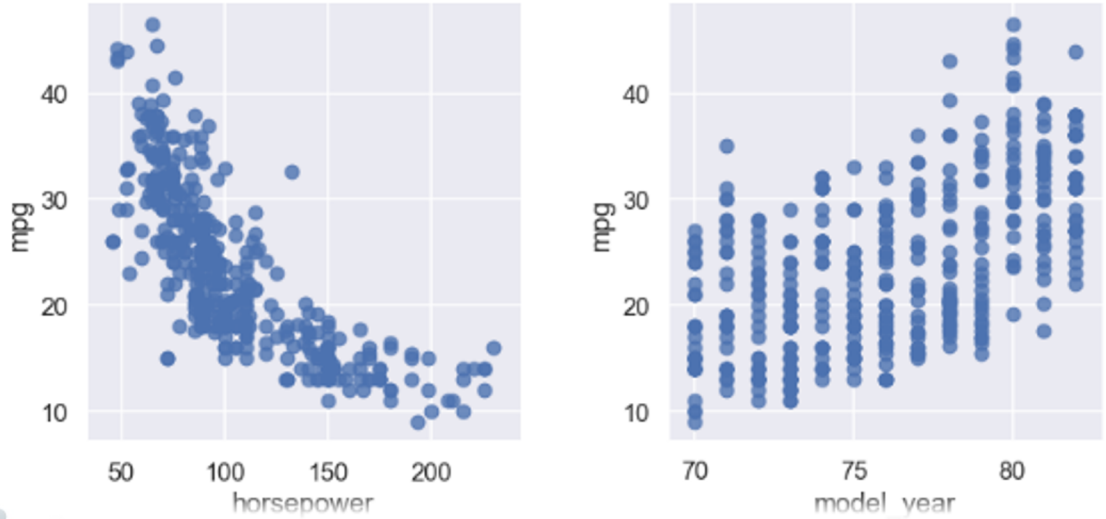
$$f_\theta(\mathbf{x}) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_p x_p$$

- **x** is a vector containing one row of input data.

MULTIVARIATE LINEAR REGRESSION

$$f_\theta(\mathbf{x}) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_p x_p$$

- Using horsepower and model year to predict mpg
  - Expect θ_1 to be negative and θ_2 to be positive. Why?



MULTIVARIATE LINEAR REGRESSION

$$f_\theta(\boldsymbol{x}) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_p x_p$$

- Many terms to write! We'll use a trick: add a column of 1s to the table:

	mpg	bias	horsepower	weight	model_year
0	18.0	1	130.0	3504	70
1	15.0	1	165.0	3693	70
2	18.0	1	150.0	3436	70

$$\boldsymbol{x} = [1, x_1, x_2, x_3]$$
$$\boldsymbol{\theta} = [\theta_0, \theta_1, \theta_2, \theta_3]$$

Bolded letters means vector or matrix.

- This means our model is: $f_\theta(\boldsymbol{x}) = \boldsymbol{\theta} \cdot \boldsymbol{x}$

MULTIVARIATE LINEAR REGRESSION

$\boldsymbol{y}$: column vector of sample points to predict

$\boldsymbol{X}$: matrix of input data $(n \times p)$

	mpg	bias	horsepower	weight	model_year
0	18.0	1	130.0	3504	70
1	15.0	1	165.0	3693	70
2	18.0	1	150.0	3436	70
...	...	...	...	...	...
395	32.0	1	84.0	2295	82
396	28.0	1	79.0	2625	82
397	31.0	1	82.0	2720	82

$\boldsymbol{X}_i$: row i of input data

$\boldsymbol{x}$: row vector of input data

$\boldsymbol{\theta}$: vector of model weights

$\hat{\boldsymbol{\theta}}$: loss-minimizing model weights

Your turn: Write the matrix expression that computes a vector with a fitted linear model's predictions for **all sample points**.

MULTIVARIATE LINEAR REGRESSION

Write the matrix expression that computes a vector with a fitted linear model's predictions for all sample points.

Prediction for one point: $\hat{\boldsymbol{\theta}} \cdot \boldsymbol{x}$

Prediction for all points: $\hat{\boldsymbol{y}} = \begin{bmatrix} \hat{\boldsymbol{\theta}} \cdot \boldsymbol{X}_1 \\ \hat{\boldsymbol{\theta}} \cdot \boldsymbol{X}_2 \\ \vdots \\ \hat{\boldsymbol{\theta}} \cdot \boldsymbol{X}_n \end{bmatrix} = \boldsymbol{X}\hat{\boldsymbol{\theta}}$

$(n \times p)(p \times 1) = (n \times 1)$

MULTIVARIATE LINEAR REGRESSION

Write the matrix expression that computes the average MSE loss for all data points (this is a scalar!).

$$L(\boldsymbol{\theta}, \boldsymbol{X}, \boldsymbol{y}) = \frac{1}{n}\sum(y_i - \boldsymbol{X_i} \cdot \boldsymbol{\theta})^2$$

$$= \frac{1}{n}||\boldsymbol{y} - \boldsymbol{X\theta}||^2$$

(Bonus points if you got this)

$$||\boldsymbol{v}||^2 = \boldsymbol{v} \cdot \boldsymbol{v} = \boldsymbol{v}^\top \boldsymbol{v}$$

Using matrix notation takes a lot of practice to get used to, but the results are worth it. Always check your dimensions!

MULTIVARIATE LINEAR REGRESSION

- How do we pick θ to minimize loss?

$$L(\boldsymbol{\theta}, \boldsymbol{X}, \boldsymbol{y}) = \frac{1}{n}\sum(y_i - \boldsymbol{X_i} \cdot \boldsymbol{\theta})^2$$

- Want to take partial derivatives for θ_0, θ_1, ...
- Instead, we'll take the **gradient** and set it equal to zero.
- This solves for all model weights at once!

$$\nabla_{\boldsymbol{\theta}} L = \begin{bmatrix} \frac{\partial}{\partial \theta_0}(L) \\ \frac{\partial}{\partial \theta_1}(L) \\ \vdots \\ \frac{\partial}{\partial \theta_p}(L) \end{bmatrix}$$

MULTIVARIATE LINEAR REGRESSION

- Skipping ahead to the answer:

$$\boldsymbol{X}^\top \boldsymbol{X}\hat{\boldsymbol{\theta}} = \boldsymbol{X}^\top \boldsymbol{y}$$

$$\hat{\boldsymbol{\theta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{y}$$

What are the matrix shapes in these expressions?

- Expression above called normal equation
- Gives a closed-form recipe for fitting linear model

MULTIVARIATE LINEAR REGRESSION

- In practice, it takes too long to compute this:

$$\hat{\boldsymbol{\theta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$$

- Inverting an (n x n) matrix takes at least $O(n^2)$ time.
  - State of the art: $O(n^{2.3})$
- Takeaway: analytic solutions are elegant but are sometimes hard to find and slow.
  - Next lecture: gradient descent

REGULARIZATION

RIDGE REGRESSION

Weight Age Temp. on Mars

Height $y \approx \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_d x_d$

For simplicity, we ignore θ_0.

How do we ensure that $\theta_k = 0$ when feature x_k is irrelevant?

Pick simplest model that explains data ➔ **generalization**

RIDGE REGRESSION

Add a penalty.

Pressure to fit data

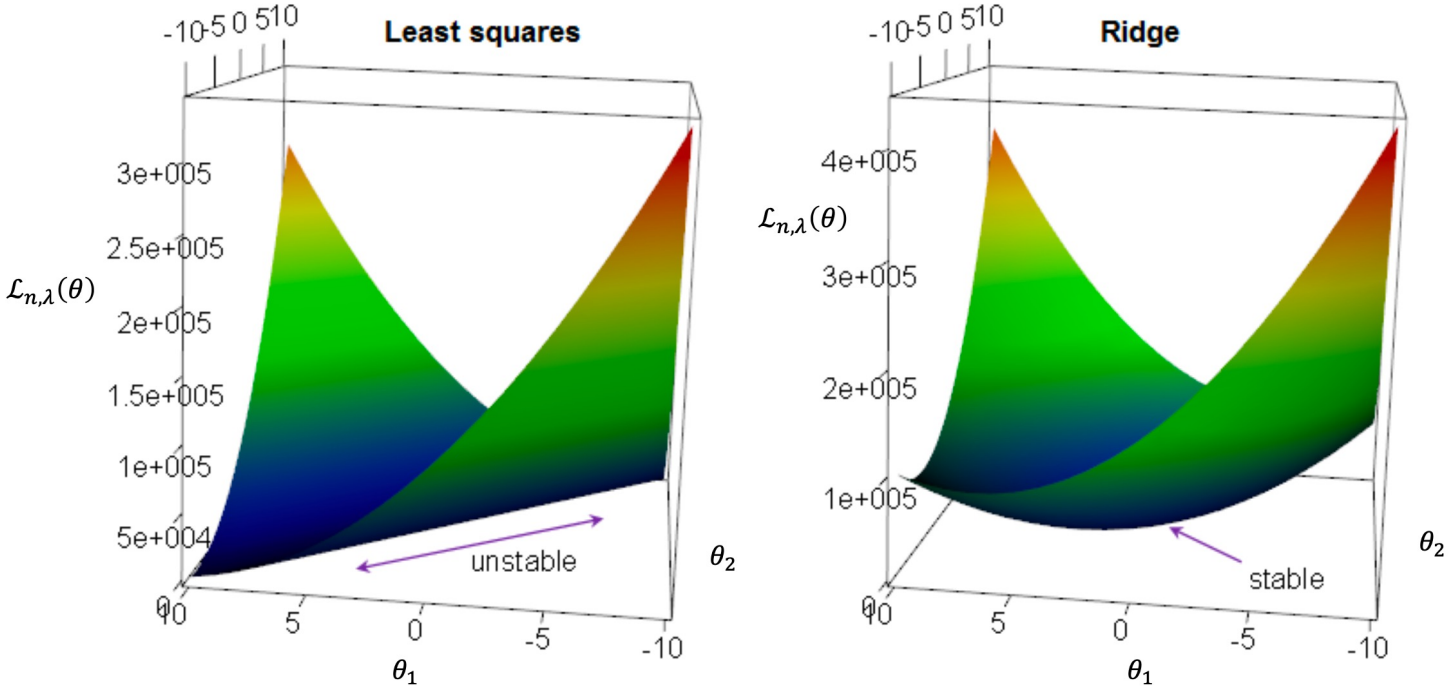
Pressure to go to zero

$$\mathcal{L}_{n,\lambda}(\theta) = \frac{1}{n}\sum_{\text{data } (x,y)} \frac{1}{2}(y - \theta^\top x)^2 + \frac{\lambda}{2}\|\theta\|^2$$

Regularization parameter $\lambda \geq 0$

Regularizer

RIDGE REGRESSION



SOLUTION

Gradient

$$\nabla \mathcal{L}_{n,\lambda}(\theta) = \lambda\theta + \frac{1}{n}(X^\top X)\theta - \frac{1}{n}X^\top Y$$

Exact Solution

$$\nabla \mathcal{L}_{n,\lambda}(\hat{\theta}) = 0 \quad \Leftrightarrow \quad \lambda\hat{\theta} + \frac{1}{n}(X^\top X)\,\hat{\theta} = \frac{1}{n}X^\top Y$$

$$\Leftrightarrow \quad \hat{\theta} = (n\lambda I + X^\top X)^{-1}X^\top Y$$

This matrix is always invertible when $\lambda > 0$.

TRAINING VS TEST LOSS

Training Loss

$$\mathcal{L}_{n,\lambda}(\theta) = \frac{1}{n}\sum_{\text{trg data } (x,y)} \frac{1}{2}(y - \theta^\top x)^2 + \frac{\lambda}{2}\|\theta\|^2$$

Test Loss

$$\mathcal{R}(\theta) = \frac{1}{n}\sum_{\text{test data } (x,y)} \frac{1}{2}(y - \theta^\top x)^2$$

REGULARIZATION EFFECT