CS5100 Foundations of Artificial Intelligence

Module 6 Lesson 10

Machine Learning 2

Some images and slides are used from CS188 UC Berkeley/AIMA with permission
All materials available at http://ai.berkely.edu / http://aima.cs.berkeley.edu

Overview

Classification

Classification using Naïve Bayes

Generalization & Overfitting

Parameter Estimation & Smoothing

Machine Learning

Machine Learning: how to acquire a model based on data or experience

- Learning parameters (e.g., probabilities)
- Learning structure
- Learning hidden concepts (e.g., clustering)

Classification Example
Spam Filter ..1

Input: one or more email(s)

Output: spam/ham label(s)

Setup:

- Get a large collection of example emails, each labeled “spam” or “ham”
- Note: someone has to hand label all this data!
- Want to learn to predict labels of new, future emails

✗	Dear Sir. First, I must solicit your confidence in this transaction, this is by virture of its nature as being utterly confidencial and top secret. …
✗	TO BE REMOVED FROM FUTURE MAILINGS, SIMPLY REPLY TO THIS MESSAGE AND PUT "REMOVE" IN THE SUBJECT. 99 MILLION EMAIL ADDRESSES FOR ONLY $99
✓	Ok, Iknow this is blatantly OT but I'm beginning to go insane. Had an old Dell Dimension XPS sitting in the corner and decided to put it to use, I know it was working pre being stuck in the corner, but when I plugged it in, hit the power nothing happened.

Classification Example
Spam Filter ..2

Features: Attributes used to make the ham / spam decision

- Words: FREE!, confidential, …
- Text Patterns: $nn, CAPS
- Non-text: SenderInContacts?, TimeOfDay …
- …

✗

Dear Sir.

First, I must solicit your confidence in this transaction, this is by virture of its nature as being utterly confidencial and top secret. …

✗

TO BE REMOVED FROM FUTURE MAILINGS, SIMPLY REPLY TO THIS MESSAGE AND PUT "REMOVE" IN THE SUBJECT.

99 MILLION EMAIL ADDRESSES
FOR ONLY $99

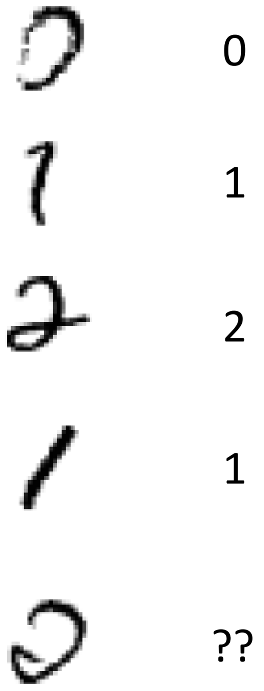
✓

Ok, Iknow this is blatantly OT but I'm beginning to go insane. Had an old Dell Dimension XPS sitting in the corner and decided to put it to use, I know it was working pre being stuck in the corner, but when I plugged it in, hit the power nothing happened.

Classification Example

Digit Recognition ..1

- Input: images / pixel grids
- Output: a digit 0-9
- Setup:
  - Get a large collection of example images, each labeled with a digit
  - Note: someone has to hand label all this data!
  - Want to learn to predict labels of new, future digit images



0

1

2

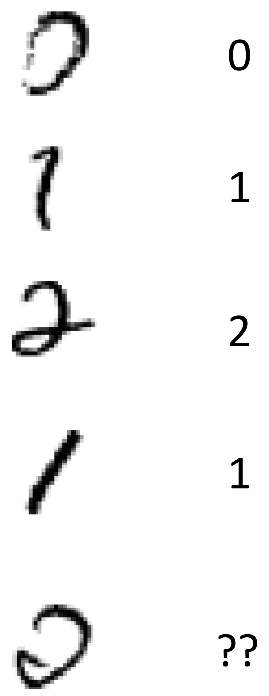
1

??

Classification Example
Digit Recognition ..2

Features: attributes used to decide on digit

- Pixels: (6,8)=ON
- Shape Patterns: NumComponents, AspectRatio, NumLoops …
- …



Other Classification Tasks

Classification: *given inputs x,*

predict labels (classes) y

Examples:

- News classification (input: news stories, classes: politics, sports, business…)
- OCR (input: images, classes: characters)
- Medical diagnosis (input: symptoms, classes: diseases)
- Automatic essay grading (input: document, classes: grades)
- Fraud detection (input: account activity, classes: fraud / no fraud)
- Customer service email routing

plus many more

Classification: important commercial technology!

Important Concepts ..1

Data: labeled instances,
e.g. emails marked spam/ham

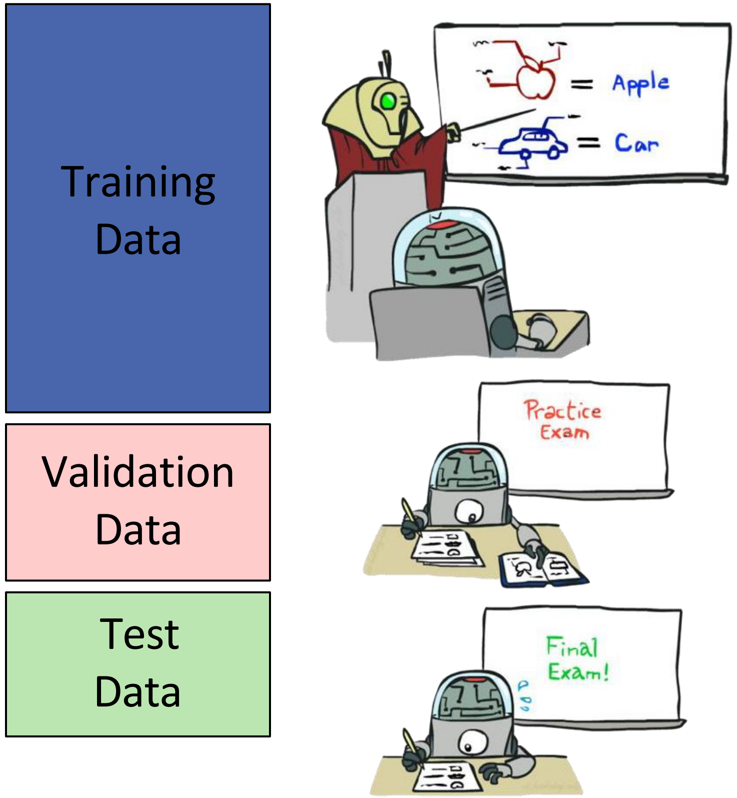
- Training set, Validation set, Test set

Features:

- attribute-value pairs which characterize each x

Experimentation cycle

- Learn parameters (e.g. model probabilities) on training set
- Tune hyperparameters on validation set
- Compute accuracy of test set
- Important: never “peek” at the test set!



Training Data

Validation Data

Test Data

Important Concepts ..2

Evaluation

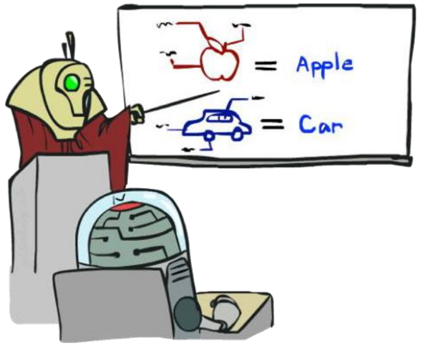
- Accuracy: fraction of instances predicted correctly

Overfitting and generalization

- Want a classifier which does well on *test* data
- Overfitting: fitting the training data very closely, but not generalizing well

Example: Chairs!

Training Data



Validation Data

Test Data

Naïve Bayes classification Intuition ..1a

Which of these is a Political story? Which is a Sports story?

- The tech giant Facebook is facing a growing backlash on Capitol Hill as more lawmakers demand that CEO Mark Zuckerberg testify about reports that Cambridge Analytica, a political data consulting firm utilized by the Trump campaign, harvested the information of up to 50 million users.
- Close to quitting tennis after four wrist surgeries in recent years, the Argentine fought to get back to the ATP Tour even as he was reduced to hitting his backhand with one hand instead of his usual two. The struggle paid off Sunday, when Del Potro ... beat top-ranked Roger Federer 6-4,7-6 (8), 7-6 (2) for the BNP Paribas Open title.

Naïve Bayes classification Intuition ..1b

Which of these is a Political story? Which is a Sports story?

- The tech giant Facebook is facing a growing backlash on Capitol Hill as more lawmakers demand that CEO Mark Zuckerberg testify about reports that Cambridge Analytica, a political data consulting firm utilized by the Trump campaign, harvested the information of up to 50 million users.
- Close to quitting tennis after four wrist surgeries in recent years, the Argentine fought to get back to the ATP Tour even as he was reduced to hitting his backhand with one hand instead of his usual two. The struggle paid off Sunday, when Del Potro ... beat top-ranked Roger Federer 6-4,7-6 (8), 7-6 (2) for the BNP Paribas Open title.

Naïve Bayes classification Intuition ..2

If we have a collection of news stories labeled as Political, Sports, Tech, Business, Crime etc.,

And if the percentage of each ‘class’ is known,

We can say:

- Probability of a story being a Political story is related to:
  - P(Word | Class): Occurrence of words in Politics stories
    - E.g. election (0.1%), Capitol Hill (0.02%), campaign (0.005%) …
  - P(Class): Percentage of each class of story in the collection
    - E.g. About 30% Political, 15% Sports, 10% Tech …

What we want: P(Class | Word)

Refresher
Bayes' Rule

Two ways to factor a joint distribution over two variables:

$$P(x,y) = P(x|y)P(y) \;\; = P(y|x)P(x)$$

Dividing by P(y) , we get:

$$P(x|y) = \frac{P(y|x)}{P(y)}P(x)$$

Or in distribution form

P(Y|X) = **P**(X|Y) **P**(Y) / **P**(X) = α**P**(X|Y) **P**(Y)

P(Class|Word) = **P**(Word | Class) * **P**(Class) / **P**(Word)

α **P**(Word | Class) * **P**(Class)

General Naïve Bayes ..1

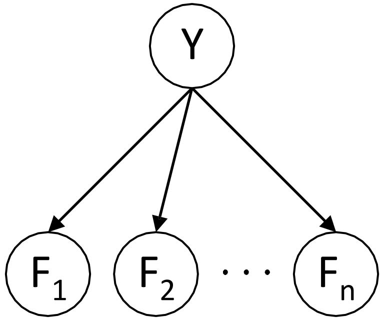
A general Naïve Bayes model:

|Y| parameters

$$\underbrace{P(\mathsf{Y}, \mathsf{F}_1 \dots \mathsf{F}_n) =}_{|Y| \times |F|^n \text{ values}} \quad \overbrace{P(\mathsf{Y})}^{|Y| \text{ parameters}} \underbrace{\prod_i P(\mathsf{F}_i|\mathsf{Y})}_{n \times |F| \times |Y| \text{ parameters}}$$

|Y| x |F|^n values

n x |F| x |Y| parameters



We only have to specify how each feature depends on the class

Total number of parameters is *linear* in n

Model is very simplistic, but often works anyway

General Naïve Bayes ..2

What do we need in order to use
Naïve Bayes?

- Estimates of local conditional probability tables
    - P(Y), the priors over labels
    - P(F$_i$|Y) for each feature (evidence variables)
    - These probabilities are collectively called the *parameters* of the model and denoted by θ
    - Up until now, we assumed these appeared by magic, but…
    - … they typically come from training data counts:
      we'll look at this soon

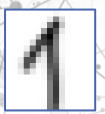
Naïve Bayes for Digits

Naïve Bayes:

Assume all features are independent

Simple digit recognition version:

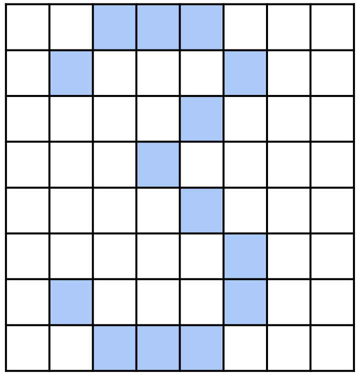
- One feature (variable) F_{ij} for each grid position <i,j>, i,j: 0..15
- Feature values are on / off, based on underlying image
- Each input maps to a feature vector, e.g.

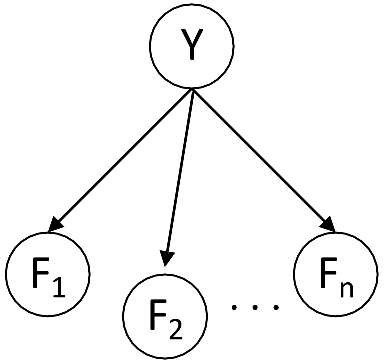
 $\rightarrow \langle F_{0,0}=0\ \ F_{0,1}=0\ \ F_{0,2}=1\ \ F_{0,3}=1\ \ F_{0,4}=0\ \ \ldots F_{15,15}=0\rangle$

Here: lots of features, each binary-valued

Naïve Bayes model: $$P(Y|F_{0,0}\ldots F_{15,15}) \propto P(Y)\prod_{i,j} P(F_{i,j}|Y)$$

What do we need to learn to classify images as digits?

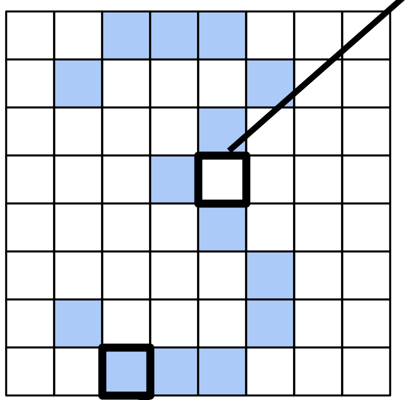




Example
Conditional Probabilities

$P(Y)$

1	0.1
2	0.1
3	0.1
4	0.1
5	0.1
6	0.1
7	0.1
8	0.1
9	0.1
0	0.1



$P(F_{5,5} = on|Y)$

1	0.05
2	0.01
3	0.90
4	0.80
5	0.90
6	0.90
7	0.25
8	0.85
9	0.60
0	0.80

$P(F_{3,1} = on|Y)$

1	0.01
2	0.05
3	0.05
4	0.30
5	0.80
6	0.90
7	0.05
8	0.60
9	0.50
0	0.80

Parameter Estimation

Estimating the distribution of a random variables like X or X | Y

Elicitation: ask a human

- Need domain experts, sophisticated ways of eliciting probabilities

Empirically: use training data (learning!)

- E.g.: for each outcome x, look at the *empirical rate* of that value:

$$P_{\text{ML}}(x) = \frac{\text{count}(x)}{\text{total samples}}$$

- This is the estimate that maximizes the *likelihood of the data*

r r b

$P_{\text{ML}}(\text{r}) = 2/3$

Naïve Bayes for Text ..1

Bag-of-words Naïve Bayes:

- Features: W_i is the word at position i
- As before: predict label (spam vs. ham) conditioned on feature variables
- As before: assume features are conditionally independent given label
- New: each W_i is identically distributed

Generative model:

$$P(Y, W_1 \dots W_n) = P(Y) \prod_i \underbrace{P(W_i|Y)}_{}$$

Word at position i, not i^{th} word in the dictionary!

Naïve Bayes for Text ..2

Bag-of-words model

- Usually, each variable gets its own conditional probability distribution P(F|Y)
- In a bag-of-words model
  - Each position is identically distributed
  - All positions share the same conditional probabilities P(W|Y)
  - Why make this assumption?
    - Called “bag-of-words” because model not dependent on word order or reordering

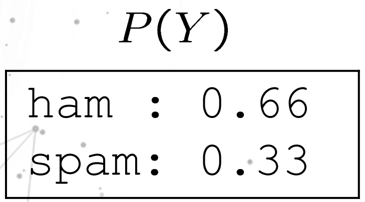
$$P(Y, W_1 \dots W_n) = P(Y) \prod_i P(W_i|Y)$$

Example
Spam Filtering

Model:

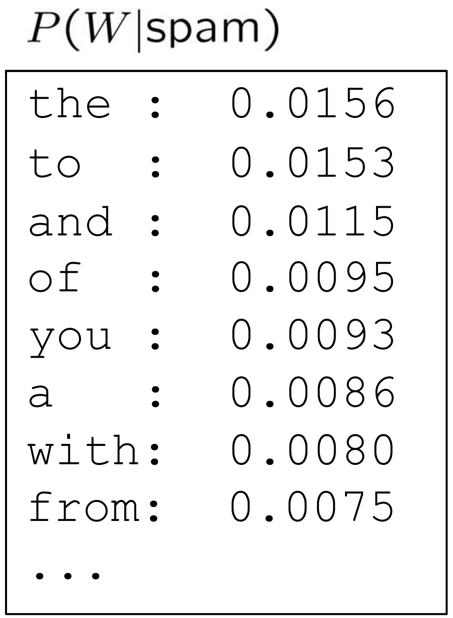
$$P(Y, W_1 \ldots W_n) = P(Y) \prod_i P(W_i|Y)$$

What are the parameters?



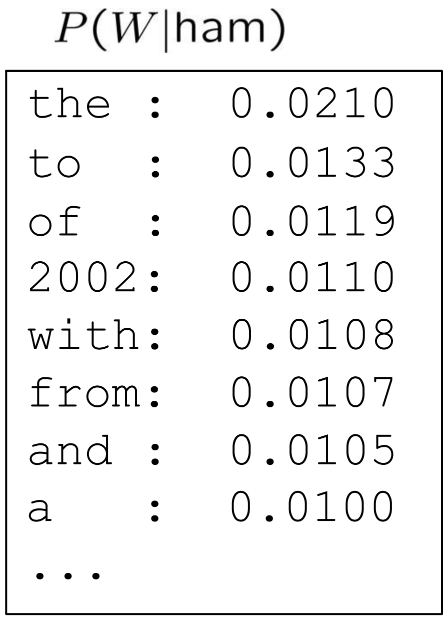
$P(Y)$

ham	:	0.66
spam	:	0.33



$P(W|\text{spam})$

the	:	0.0156
to	:	0.0153
and	:	0.0115
of	:	0.0095
you	:	0.0093
a	:	0.0086
with	:	0.0080
from	:	0.0075
...		



$P(W|\text{ham})$

the	:	0.0210
to	:	0.0133
of	:	0.0119
2002	:	0.0110
with	:	0.0108
from	:	0.0107
and	:	0.0105
a	:	0.0100
...		

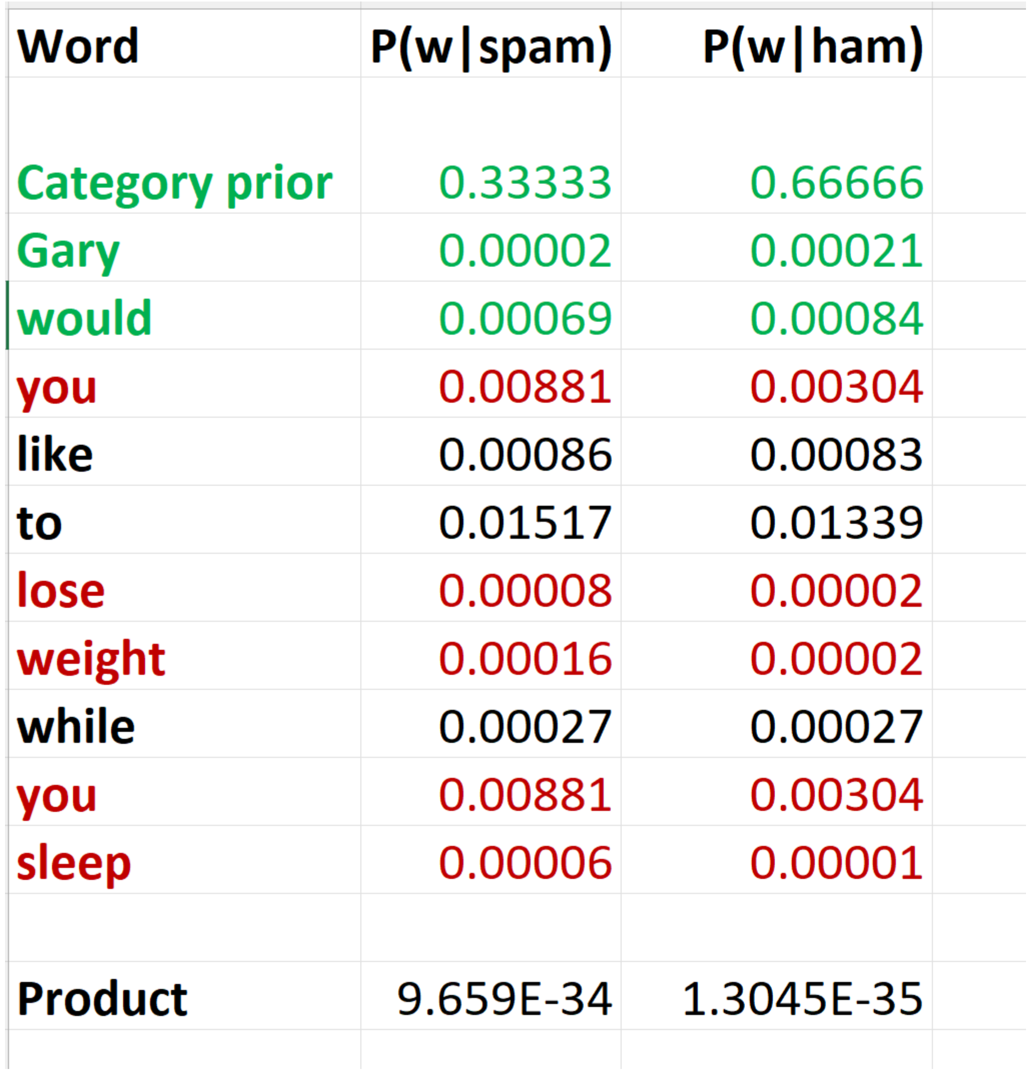
Where do these tables come from?

Counts from examples!

Spam Example

$$P(Y, W_1 \dots W_n) = P(Y) \prod_i P(W_i|Y)$$

$$= P(Outcome) * \prod_i P(Event_i|Outcome)$$



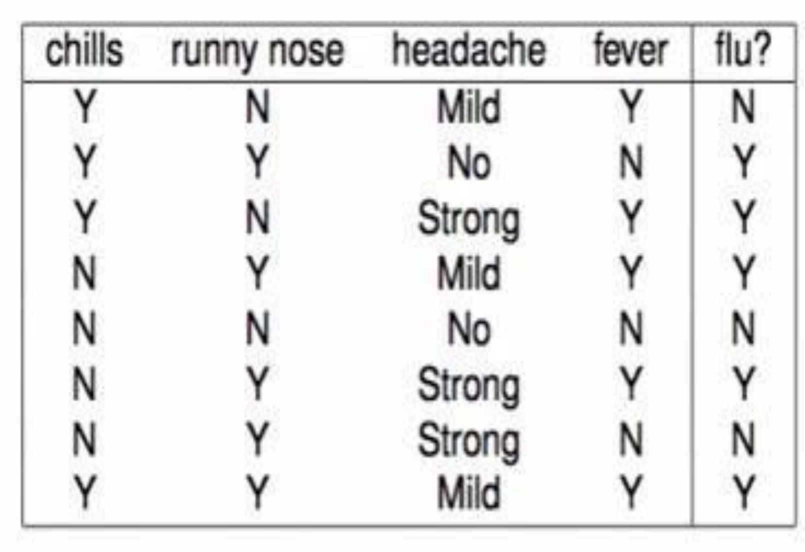
Word	P(w|spam)	P(w|ham)
Category prior	0.33333	0.66666
Gary	0.00002	0.00021
would	0.00069	0.00084
you	0.00881	0.00304
like	0.00086	0.00083
to	0.01517	0.01339
lose	0.00008	0.00002
weight	0.00016	0.00002
while	0.00027	0.00027
you	0.00881	0.00304
sleep	0.00006	0.00001
Product	9.659E-34	1.3045E-35

Naïve Bayes Quick Check ..1

P(Outcome | Events) =

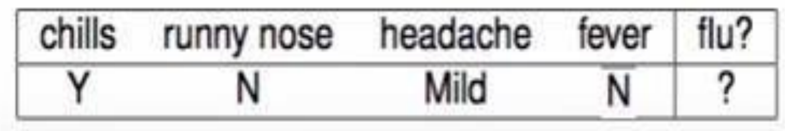
$$P(\text{Outcome}) * \frac{\varsigma_i\ P(Event_i\ |Outcome)}{\varsigma_i\ P(Eventi)}$$

Training Data



chills	runny nose	headache	fever	flu?
Y	N	Mild	Y	N
Y	Y	No	N	Y
Y	N	Strong	Y	Y
N	Y	Mild	Y	Y
N	N	No	N	N
N	Y	Strong	Y	Y
N	Y	Strong	N	N
Y	Y	Mild	Y	Y

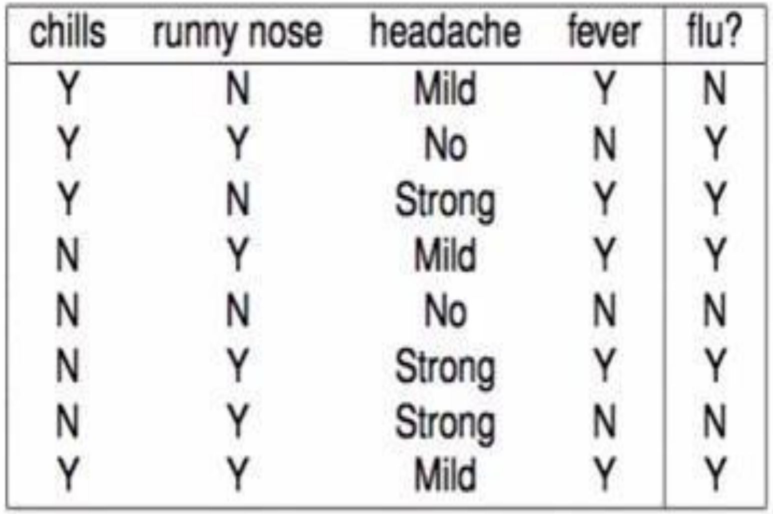
Test Data



chills	runny nose	headache	fever	flu?
Y	N	Mild	N	?

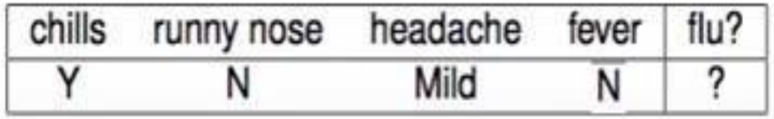
Naïve Bayes Quick Check ..2

Training Data



chills	runny nose	headache	fever	flu?
Y	N	Mild	Y	N
Y	Y	No	N	Y
Y	N	Strong	Y	Y
N	Y	Mild	Y	Y
N	N	No	N	N
N	Y	Strong	Y	Y
N	Y	Strong	N	N
Y	Y	Mild	Y	Y

Test Data



chills	runny nose	headache	fever	flu?
Y	N	Mild	N	?

P(Outcome | Events) =

$$P(Outcome) * \frac{\varsigma_i\ P(Event_i\ |Outcome)}{\varsigma_i\ P(Eventi)}$$

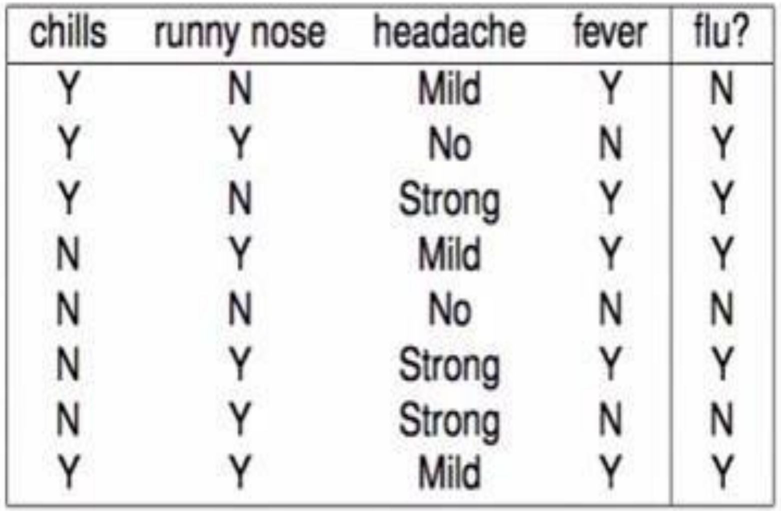
- P(flu = Y) = 5/8 = 0.625
- P(chills=Y | flu) = 3/5 = 0.6
- P(runnynose=N | flu) = 1/5 = 0.2
- P(headache=mild | flu) = 2/5 = 0.4
- P(fever=N | flu) = 1/5 = 0.2
- P(chills=Y) = 4/8 = 0.5
- P(runnynose=N) = 3/8 = 0.375
- P(headache=mild) = 3/8 = 0.375
- P(fever=N) = 3/8 = 0.375

- P(flu = Y |Events) =

$$\frac{0.625 * 0.6 * 0.2 * 0.4 * 0.2}{0.5 * 0.375 * 0.375 * 0.375} = 0.228$$

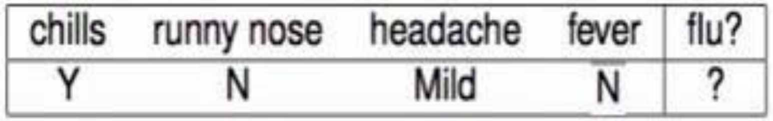
Naïve Bayes Quick Check ..3

Training Data



chills	runny nose	headache	fever	flu?
Y	N	Mild	Y	N
Y	Y	No	N	Y
Y	N	Strong	Y	Y
N	Y	Mild	Y	Y
N	N	No	N	N
N	Y	Strong	Y	Y
N	Y	Strong	N	N
Y	Y	Mild	Y	Y

Test Data



chills	runny nose	headache	fever	flu?
Y	N	Mild	N	?

P(Outcome | Events) =

$$P(\text{Outcome}) * \frac{\varsigma_i \; P(Event_i \,|Outcome)}{\varsigma_i \; P(Eventi)}$$

- P(flu = N) = 3/8 = 0.375
- P(chills=Y | flu=N) = 0.33
- P(runnynose=N | flu=N) = 0.66
- P(headache=mild | flu=N) = 0.33
- P(fever=N | flu=N) = 0.66
- P(chills=Y) = 4/8 = 0.5
- P(runnynose=N) = 3/8 = 0.375
- P(headache=mild) = 3/8 = 0.375
- P(fever=N) = 3/8 = 0.375

- P(flu = N |Events) =

$$\frac{0.375 \ast 0.33 \ast 0.66 \ast 0.33 \ast 0.66}{0.5 \ast 0.375 \ast 0.375 \ast 0.375} = 0.675$$

A simplification

Denominator d = ς P(Event$_i$) is common to both (all)

Usually we’re trying to get relative probabilities of outcomes

So denominator may be dropped

P(flu = Y |Events) = 0.006 / d

P(flu = N |Events) = 0.0178 / d

So: No flu!

Naïve Bayes: Advantages

Simple conceptual model

Works well in a lot of cases

Fast and easy to compute:

- Determine probabilities
  Perform some multiplications
- But not great at estimating probabilities

Easy to retrain (update) when you get additional training data

Naïve Bayes & scikit-learn

Three variants

based on different methods to compute $P(E_i|O)$

Gaussian NB

features assumed to be in normal distribution

Multinomial NB

used for text problems, where data is represented as word count or tf-idf vectors; incorporates smoothing*

Bernoulli NB

may be multiple features, but all features assumed to be binary-valued

Example Overfitting

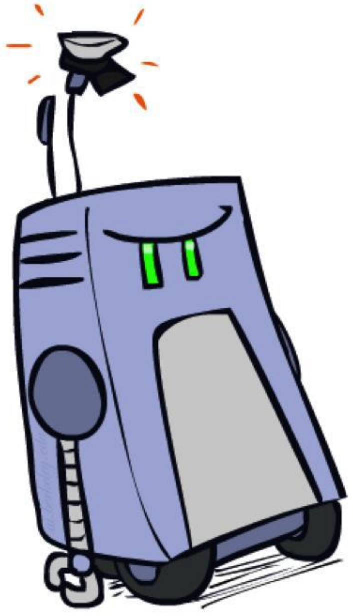
Raw probabilities alone don’t affect the posteriors; relative probabilities (odds ratios) do:

$$\frac{P(W|\text{ham})}{P(W|\text{spam})}$$

```
south-west  :  inf
nation      :  inf
morally     :  inf
nicely      :  inf
extent      :  inf
seriously   :  inf
...
```

$$\frac{P(W|\text{spam})}{P(W|\text{ham})}$$

```
screens     :  inf
minute      :  inf
guaranteed  :  inf
$205.00     :  inf
delivery    :  inf
signature   :  inf
...
```



What went wrong here?

Generalization and Overfitting ..1

Relative frequency parameters will overfit the training data!

- Just because we never saw a 3 with pixel (15,15) **on** during training doesn’t mean we won’t see it at test time
- Unlikely that every occurrence of “minute” is 100% spam
- Unlikely that every occurrence of “seriously” is 100% ham
- What about all the words that don’t occur in the training set at all?
- In general, we can’t go around giving unseen events zero probability

Generalization and Overfitting ..2

- As an extreme case, imagine using the entire email as the only feature
  - Would get the training data perfect (if deterministic labeling)
  - Wouldn't *generalize* at all
  - Just making the bag-of-words assumption gives us some generalization, but is not enough

- To generalize better: we need to smooth or regularize the estimates

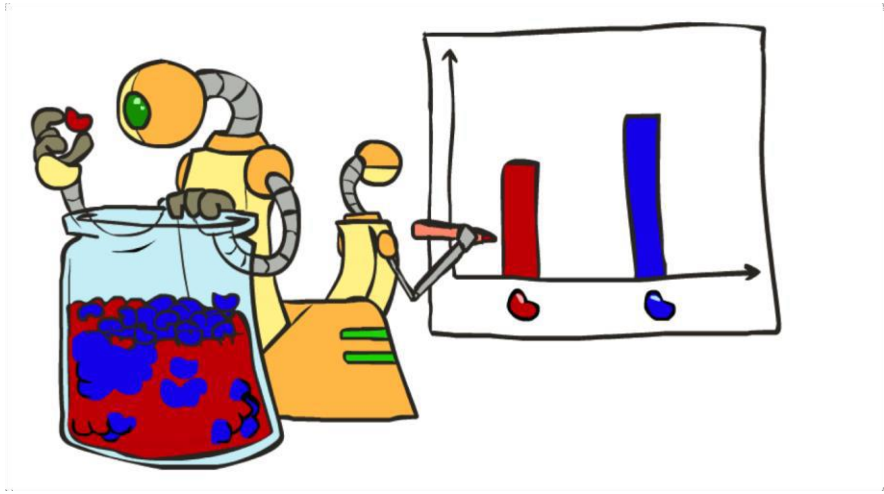
Estimation & Smoothing

Problems with maximum likelihood estimates:

- If I flip a coin once, and it comes up heads, what is the estimate for P(heads)?
- What if we have 8 heads after 10 flips?
- What if we have 8 million heads after 10 million flips?

Basic idea behind solution:

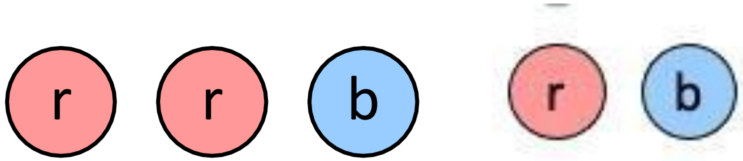
- We have a prior expectation about parameters – e.g. P(heads) here
- Given only little evidence, we should skew towards our prior expectation
- Given a lot of evidence, we should listen to the data



Laplace Smoothing ..1

Laplace's estimate:

Pretend you saw every outcome once more than you actually did



$$P_{LAP}(x) = \frac{c(x)+1}{\sum_x [c(x)+1]}$$

$$= \frac{c(x)+1}{N+|X|}$$

$$P_{ML}(X) =$$

$$P_{LAP}(X) =$$

Laplace Smoothing ..2

Laplace's estimate (extended):

Pretend you saw every outcome k extra times

$$P_{LAP,k}(x) = \frac{c(x) + k}{N + k|X|}$$

What's Laplace with k = 0?

r r b

$$P_{LAP,0}(X) =$$

$$P_{LAP,1}(X) =$$

$$P_{LAP,100}(X) =$$

Real NB Smoothing

For real classification problems, smoothing is critical

New odds ratios:

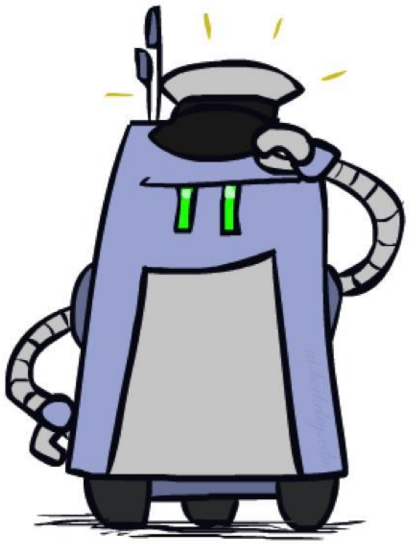
$$\frac{P(W|\text{ham})}{P(W|\text{spam})}$$

helvetica	:	11.4
seems	:	10.8
group	:	10.2
ago	:	8.4
areas	:	8.3
...		

$$\frac{P(W|\text{spam})}{P(W|\text{ham})}$$

verdana	:	28.8
Credit	:	28.4
ORDER	:	27.2
<FONT>	:	26.9
money	:	26.5
...		

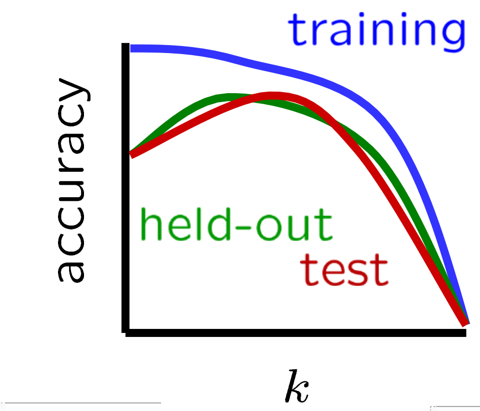
Do these make more sense?

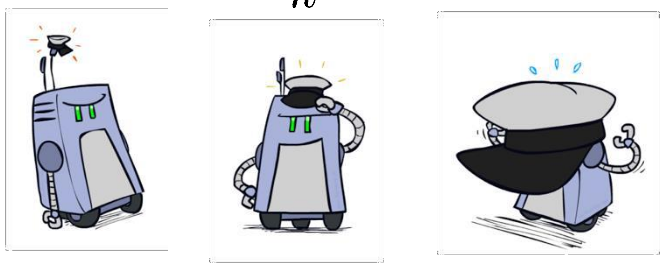


Tuning on Held-Out Data

- Now we’ve got two kinds of unknowns
  - Parameters: the probabilities P(X|Y), P(Y)
  - Hyperparameters: e.g. the amount / type of smoothing to do, k, α

- What should we learn where?
  - Learn parameters from training data
  - Tune hyperparameters on different data
    - Why?
  - For each value of the hyperparameters, train and test on the held-out data
  - Choose the best value and do a final test on the test data





Baselines

First step: get a baseline

- Baselines are very simple “straw man” procedures
- Help determine how hard the task is
- Help know what a “good” accuracy is

Weak baseline: most frequent label classifier

- Gives all test instances whatever label was most common in the training set
- E.g. for spam filtering, might label everything as ham
- Accuracy might be very high if the problem is skewed
- E.g. calling everything “ham” gets 66%, so a classifier that gets 70% isn’t very good…

For real research, usually use previous work as a (strong) baseline

Examples of Ham Looking like Spam

Dear GlobalSCAPE Customer, GlobalSCAPE has partnered with ScanSoft to offer you the latest version of OmniPage Pro, for just $99.99* - the regular list price is $499! The most common question we've received about this offer is - Is this genuine? We would like to assure you that this offer is authorized by ScanSoft, is genuine and valid. You can get the . . . . . .

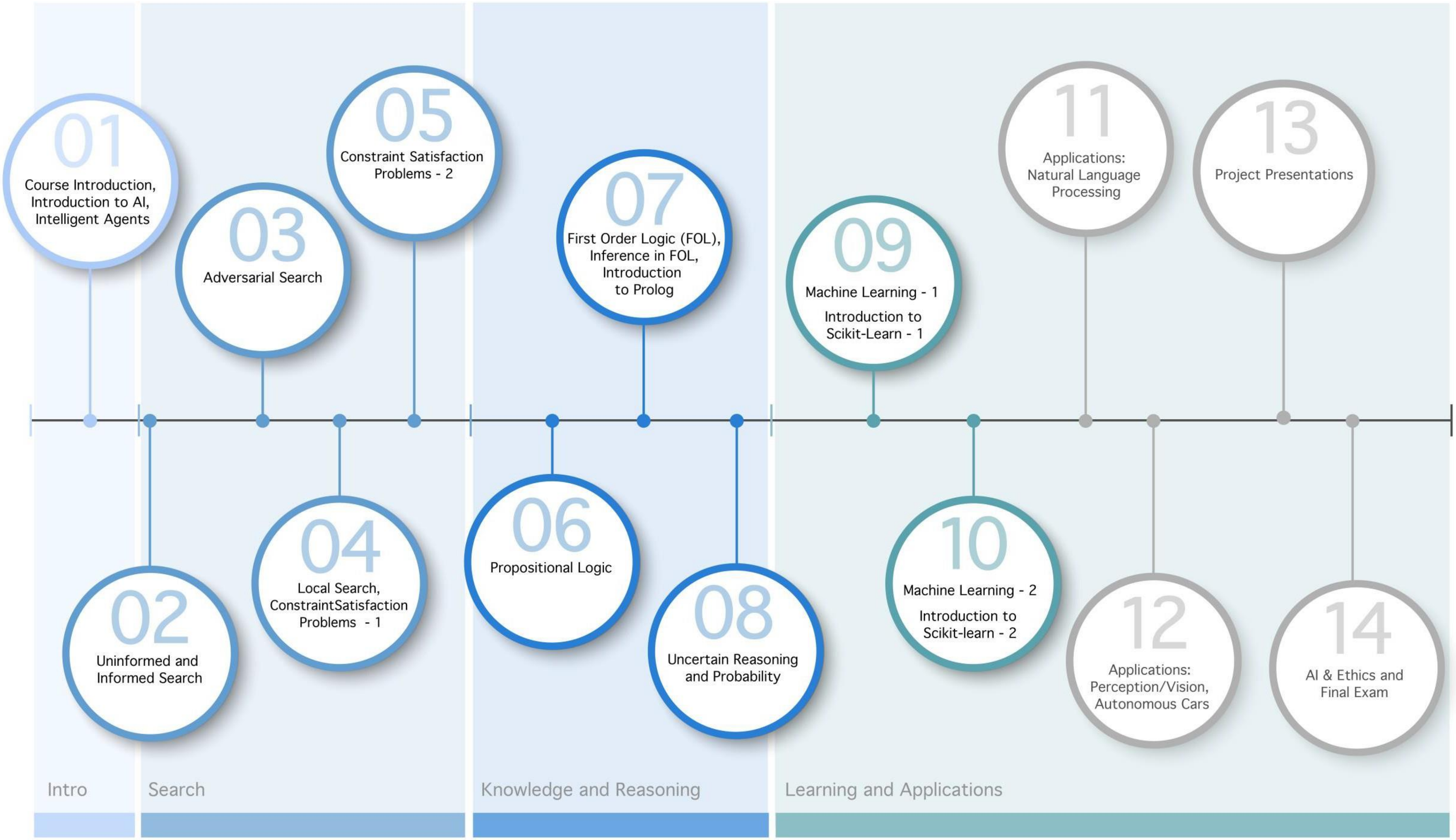
To receive your $30 Amazon.com promotional certificate, click through to http://www.amazon.com/apparel and see the prominent link for the $30 offer. All details are there. We hope you enjoyed receiving this message. However, if you'd rather not receive future e-mails announcing new store launches, please click . . .

What to Do About Errors?

Need more features– words aren’t enough!

- Have you emailed the sender before?
- Have 1K other people just gotten the same email?
- Is the sending information consistent?
- Is the email in ALL CAPS?
- Do inline URLs point where they say they point?
- Does the email address you by (your) name?

Can add these information sources as new variables in the NB model

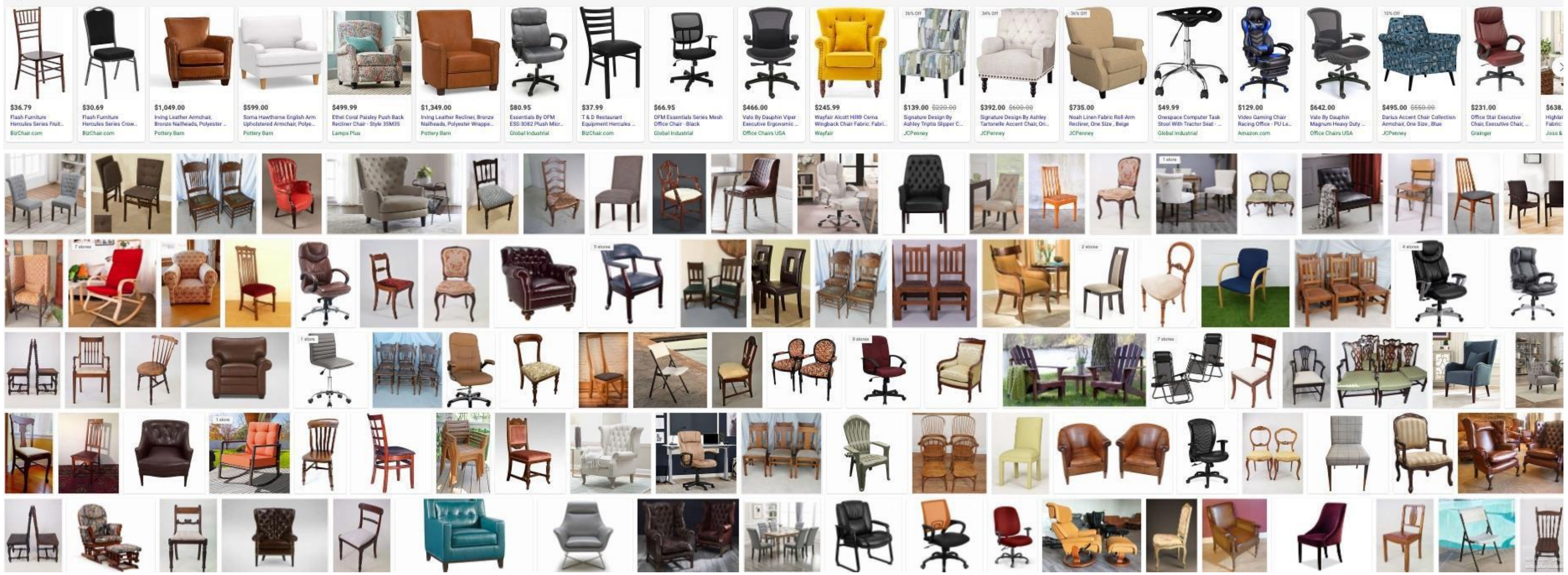

01
Course Introduction, Introduction to AI, Intelligent Agents
02
Uninformed and Informed Search
03
Adversarial Search
04
Local Search, ConstraintSatisfaction Problems - 1
05
Constraint Satisfaction Problems - 2
06
Propositional Logic
07
First Order Logic (FOL), Inference in FOL, Introduction to Prolog
08
Uncertain Reasoning and Probability
09
Machine Learning - 1
Introduction to Scikit-Learn - 1
10
Machine Learning - 2
Introduction to Scikit-learn - 2
11
Applications: Natural Language Processing
12
Applications: Perception/Vision, Autonomous Cars
13
Project Presentations
14
AI & Ethics and Final Exam
Intro
Search
Knowledge and Reasoning
Learning and Applications

Extra: ML & Chairs

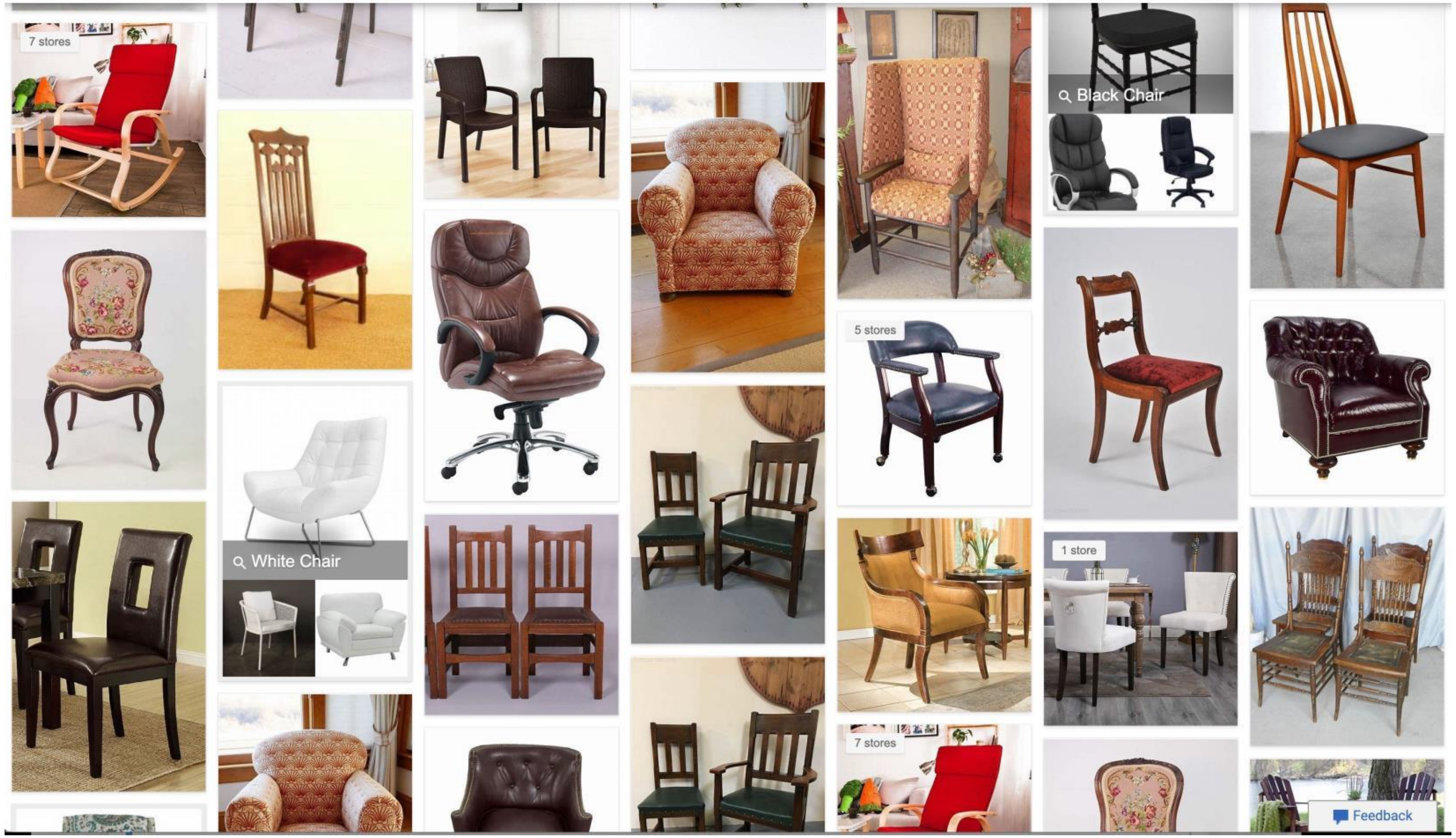
The Chair Example

https://www.westelm.com/products/windsor-dining-Chair -h5047

What “is” a Chair ?


$36.79
Flash Furniture Hercules Series Fruit...
BizChair.com
$30.69
Flash Furniture Hercules Series Crow...
BizChair.com
$1,049.00
Irving Leather Armchair, Bronze Nailheads, Polyester ...
Pottery Barn
$599.00
Soma Hawthorne English Arm Upholstered Armchair, Polye...
Pottery Barn
$499.99
Ethel Coral Paisley Push Back Recliner Chair - Style 35M35
Lamps Plus
$1,349.00
Irving Leather Recliner, Bronze Nailheads, Polyester Wrappe...
Pottery Barn
$80.95
Essentials By OFM ESS-3082 Plush Micr...
Global Industrial
$37.99
T & D Restaurant Equipment Hercules ...
BizChair.com
$66.95
OFM Essentials Series Mesh Office Chair - Black
Global Industrial
$466.00
Valo By Dauphin Viper Executive Ergonomic ...
Office Chairs USA
$245.99
Wayfair Alcott Hill® Cerna Wingback Chair Fabric: Fabri...
Wayfair
$139.00 ~~$220.00~~
Signature Design By Ashley Triptis Slipper C...
JCPenney
$392.00 ~~$600.00~~
Signature Design By Ashley Tartonelle Accent Chair, On...
JCPenney
$735.00
Noah Linen Fabric Roll-Arm Recliner, One Size , Beige
JCPenney
$49.99
Onespace Computer Task Stool With Tractor Seat - ...
Global Industrial
$129.00
Video Gaming Chair Racing Office - PU Le...
Amazon.com
$642.00
Valo By Dauphin Magnum Heavy Duty ...
Office Chairs USA
$495.00 ~~$550.00~~
Darius Accent Chair Collection Armchair, One Size , Blue
JCPenney
$231.00
Office Star Executive Chair, Executive Chair, ...
Grainger
$638.
Highlar
Fabric:
Joss &
1 store
7 stores
5 stores
2 stores
4 stores

How would you describe a Chair ? .. 1

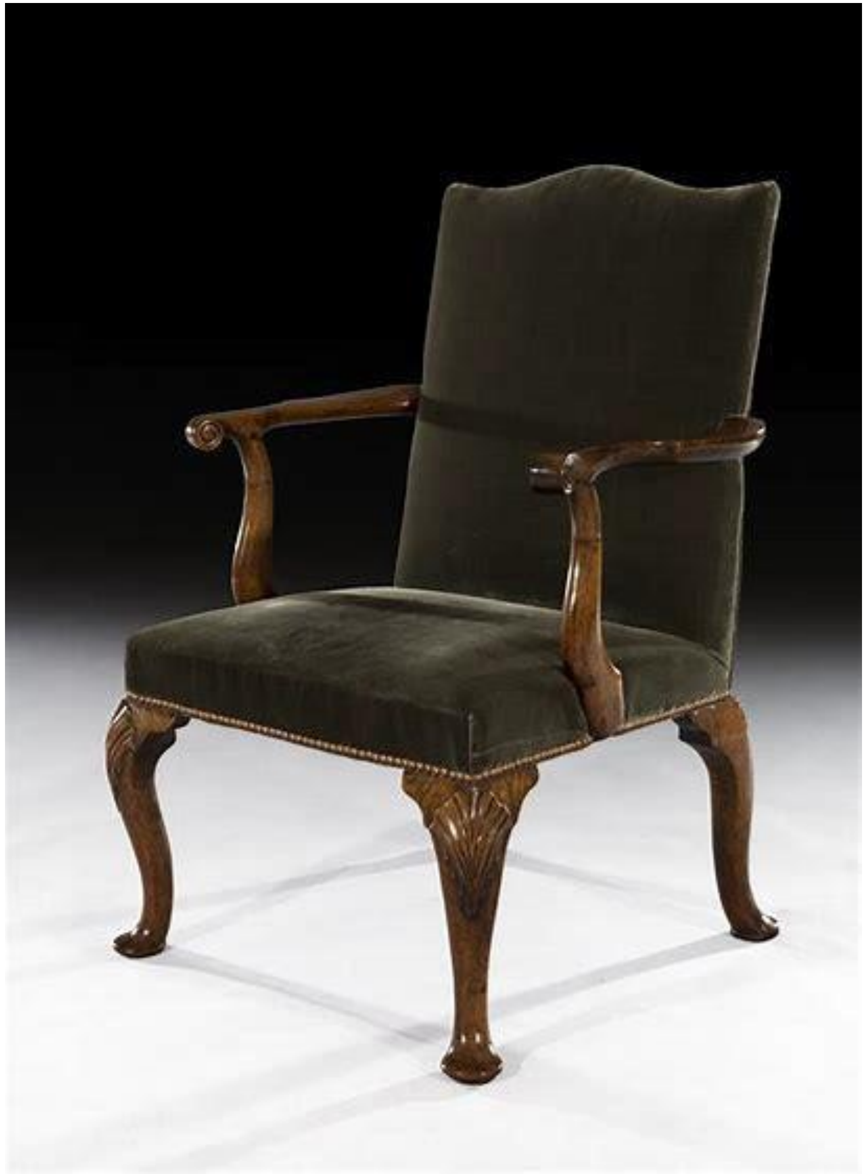


How would you **model** a chair ? .. 2

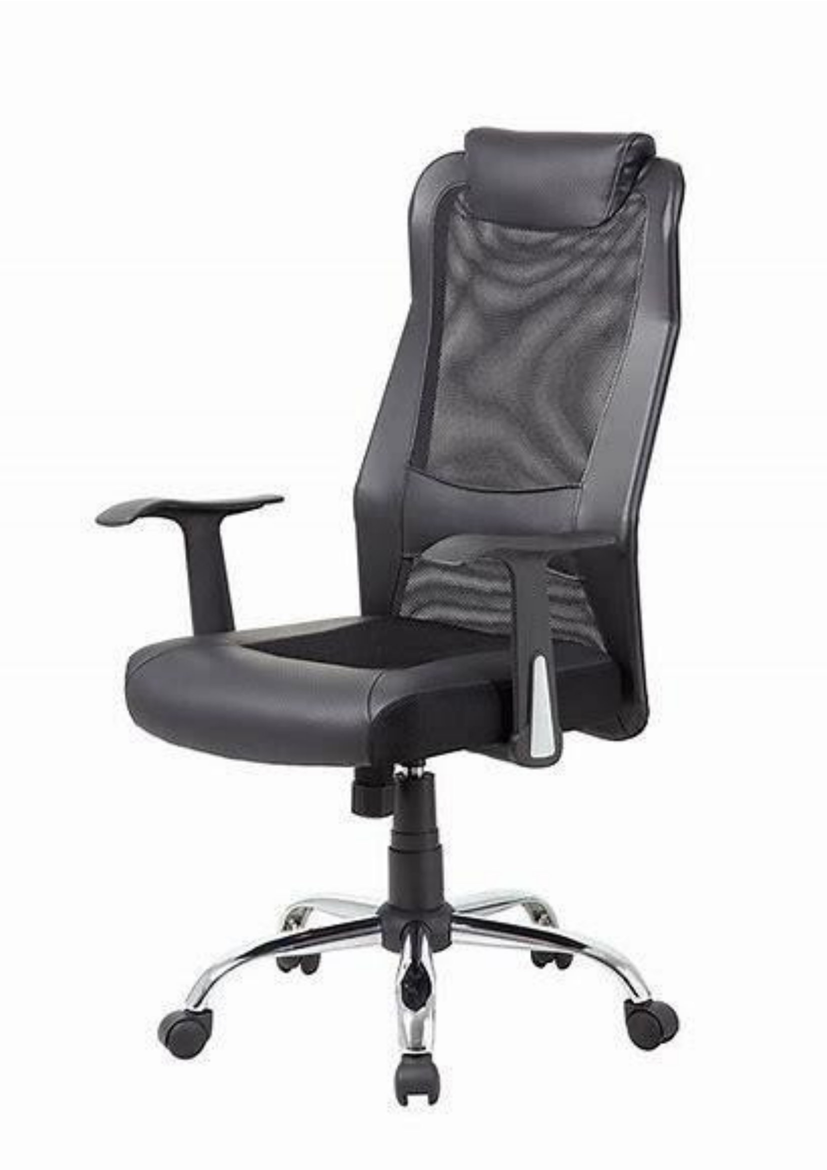
Chair 0

All images of Chairs from bing.com searches on Chairs, stools, loveseats, beanbags,…

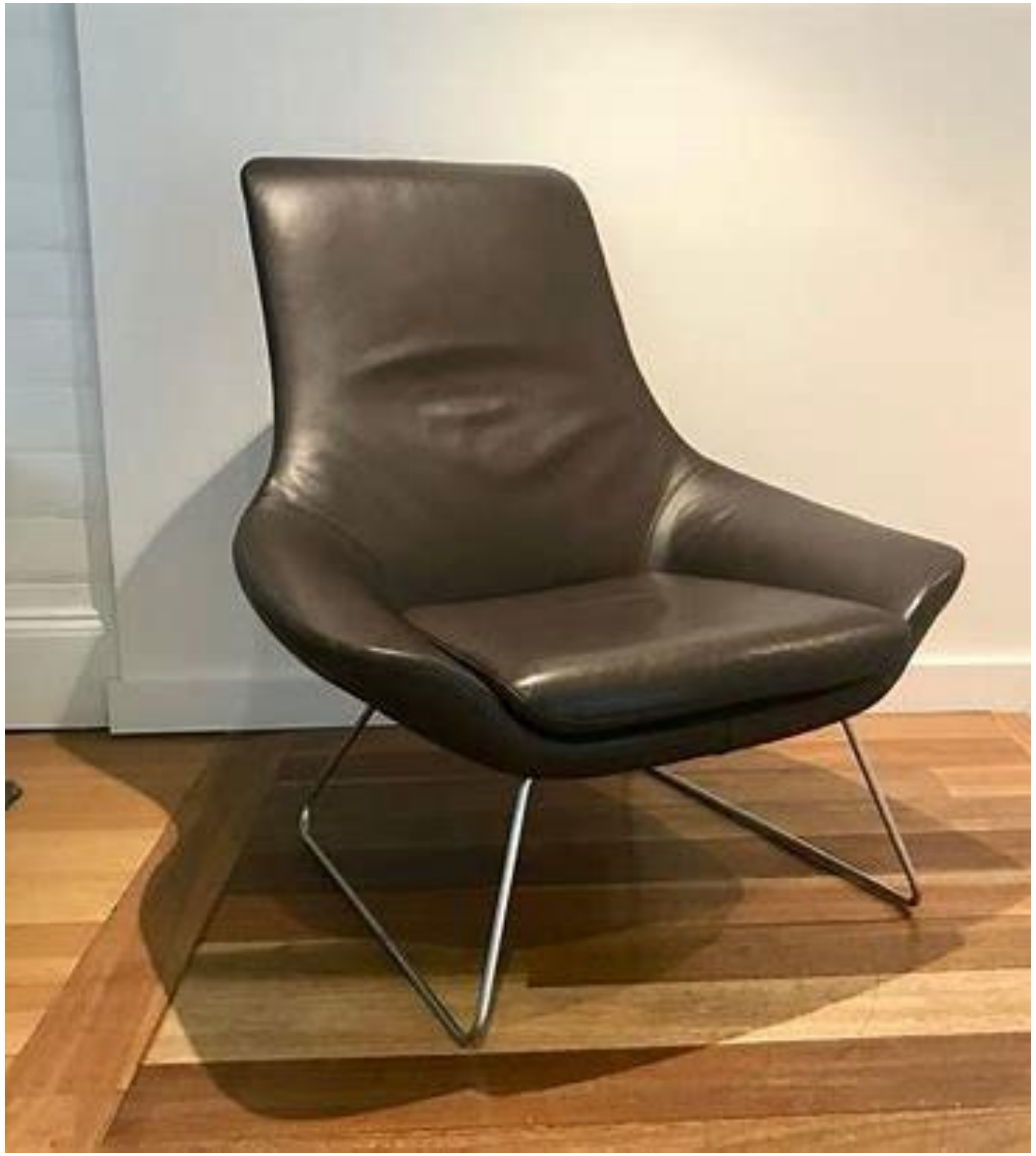
Chair 1



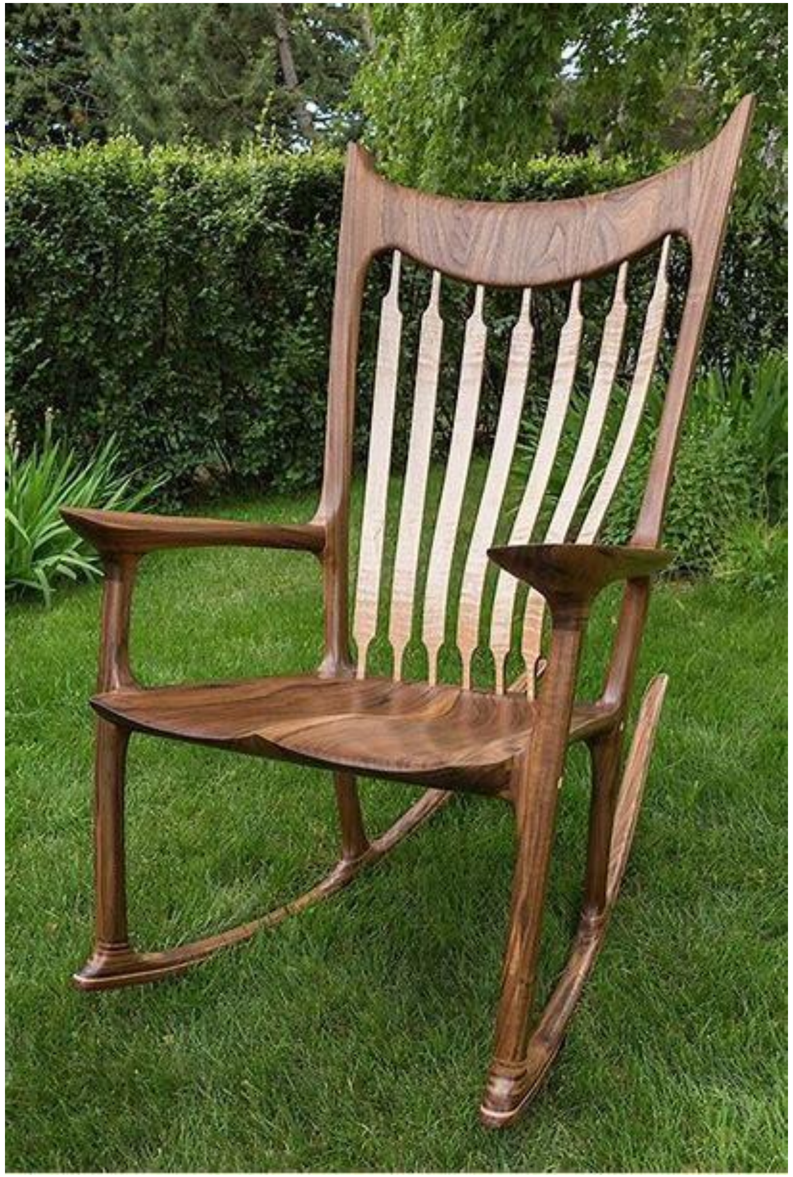
Chair 2



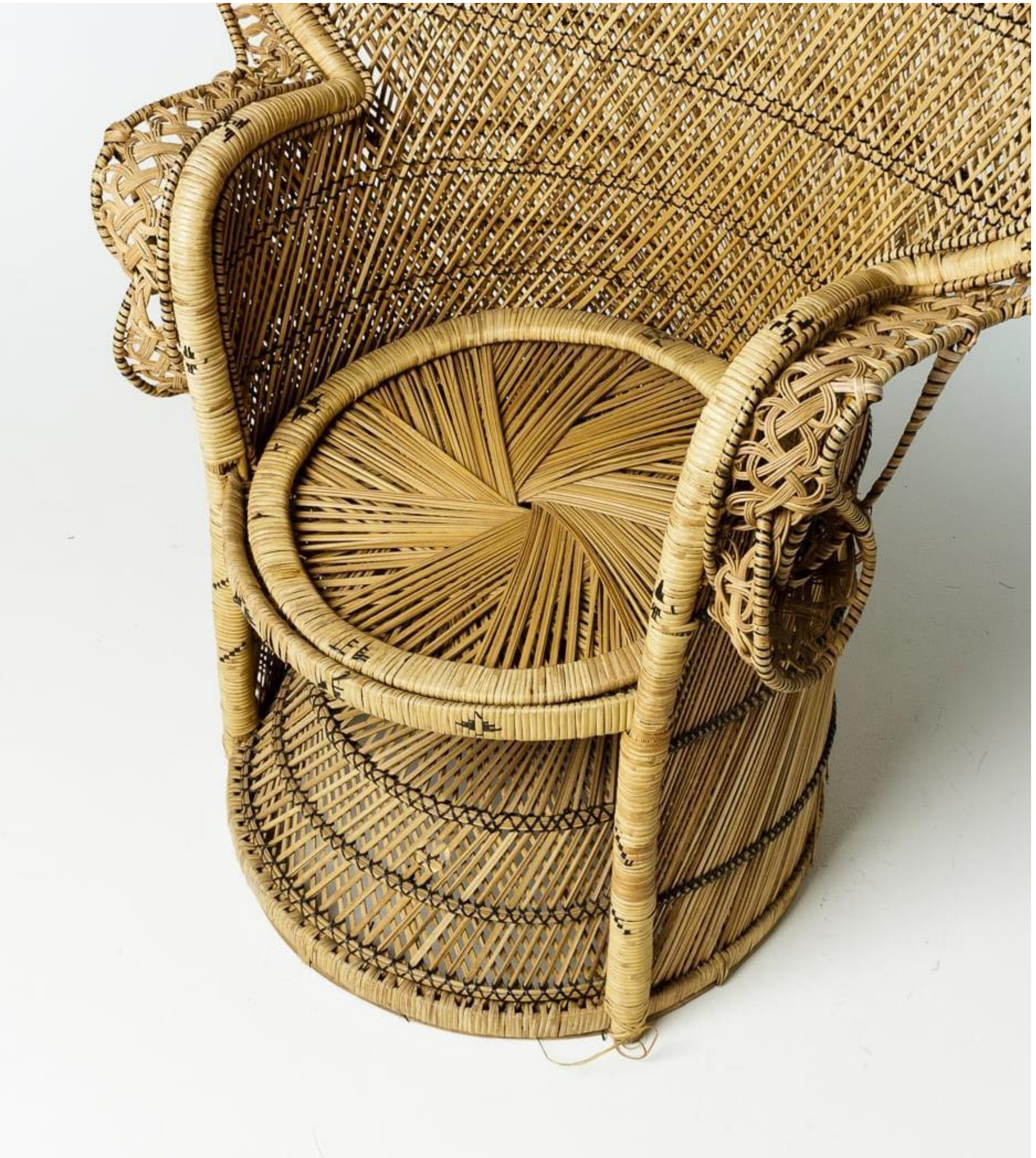
Chair 3



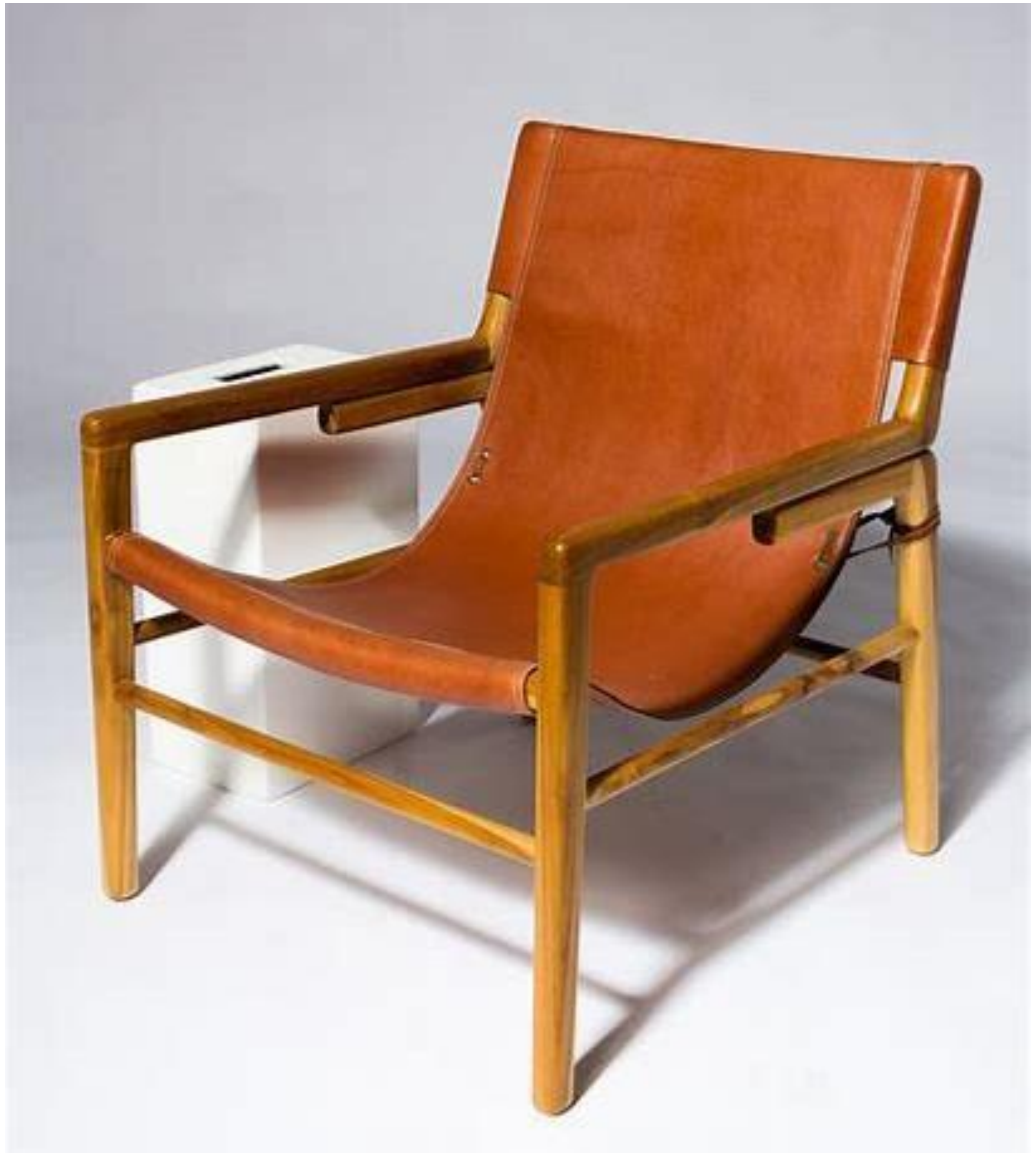
Chair 4



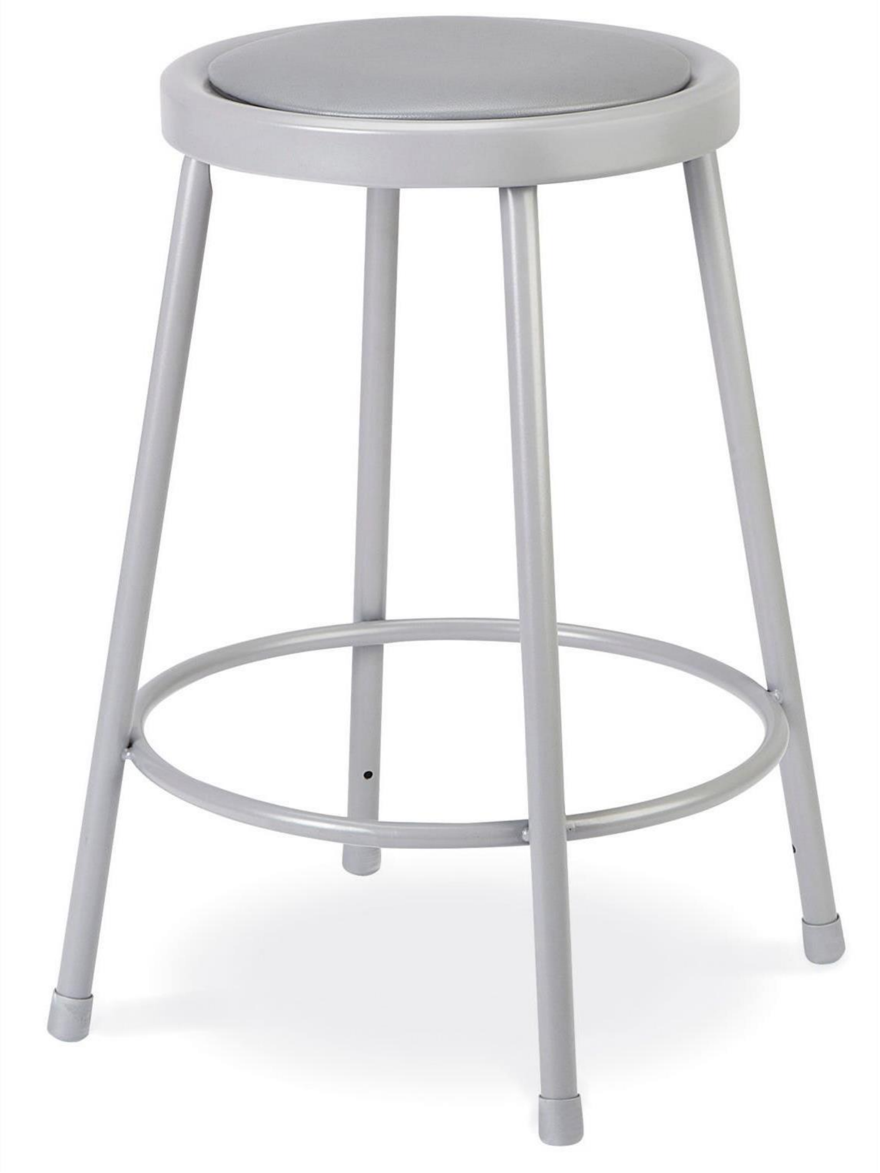
Chair 5



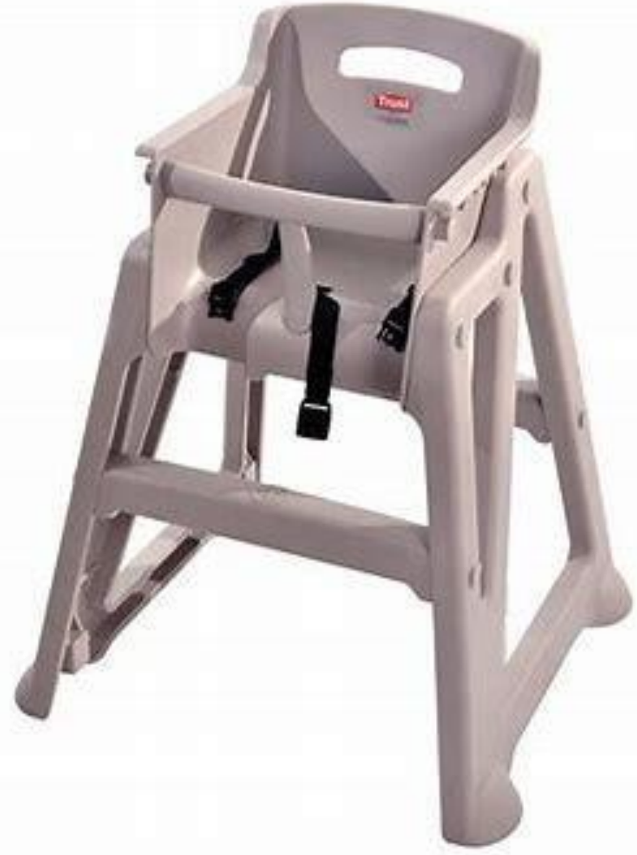
Chair 6



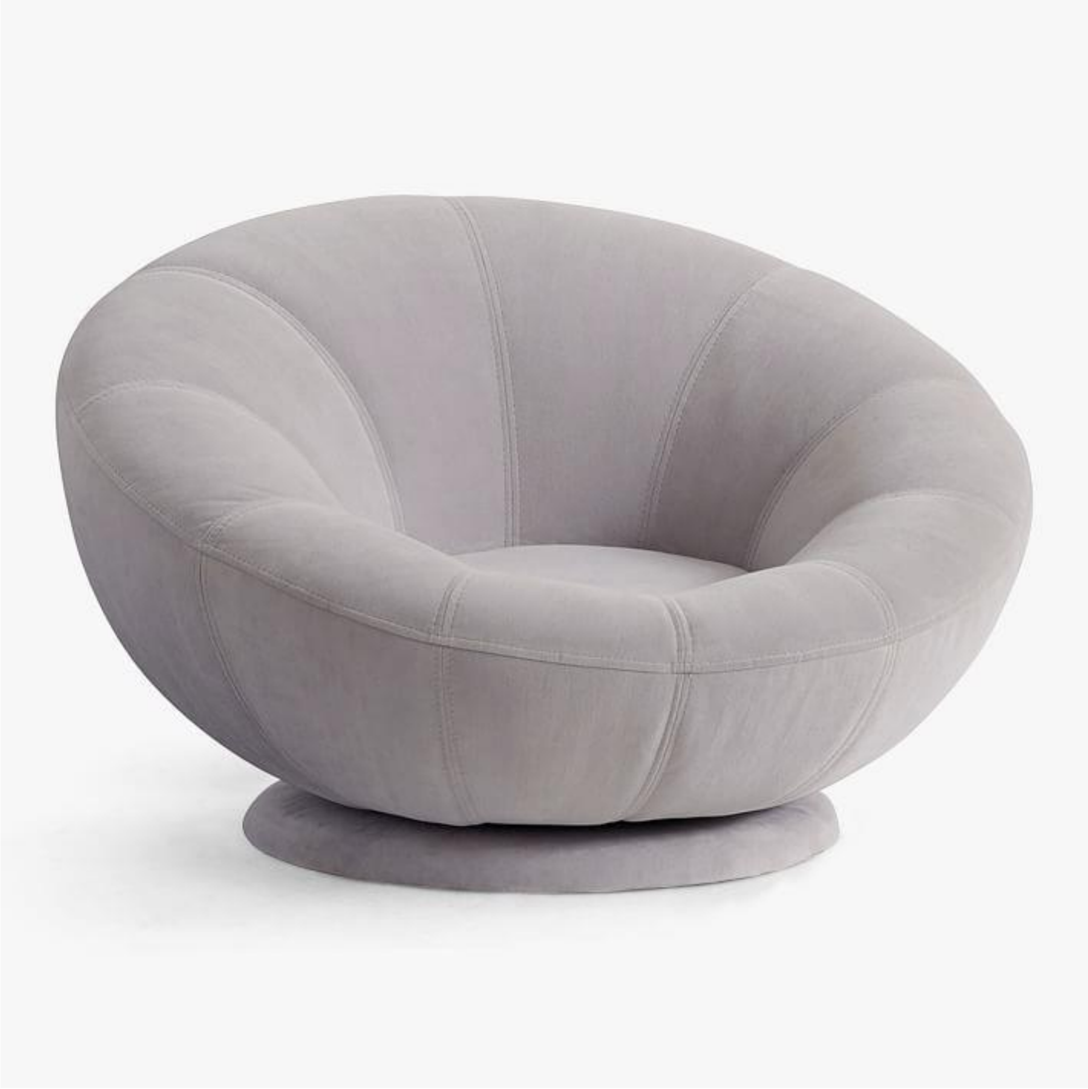
Chair ?7



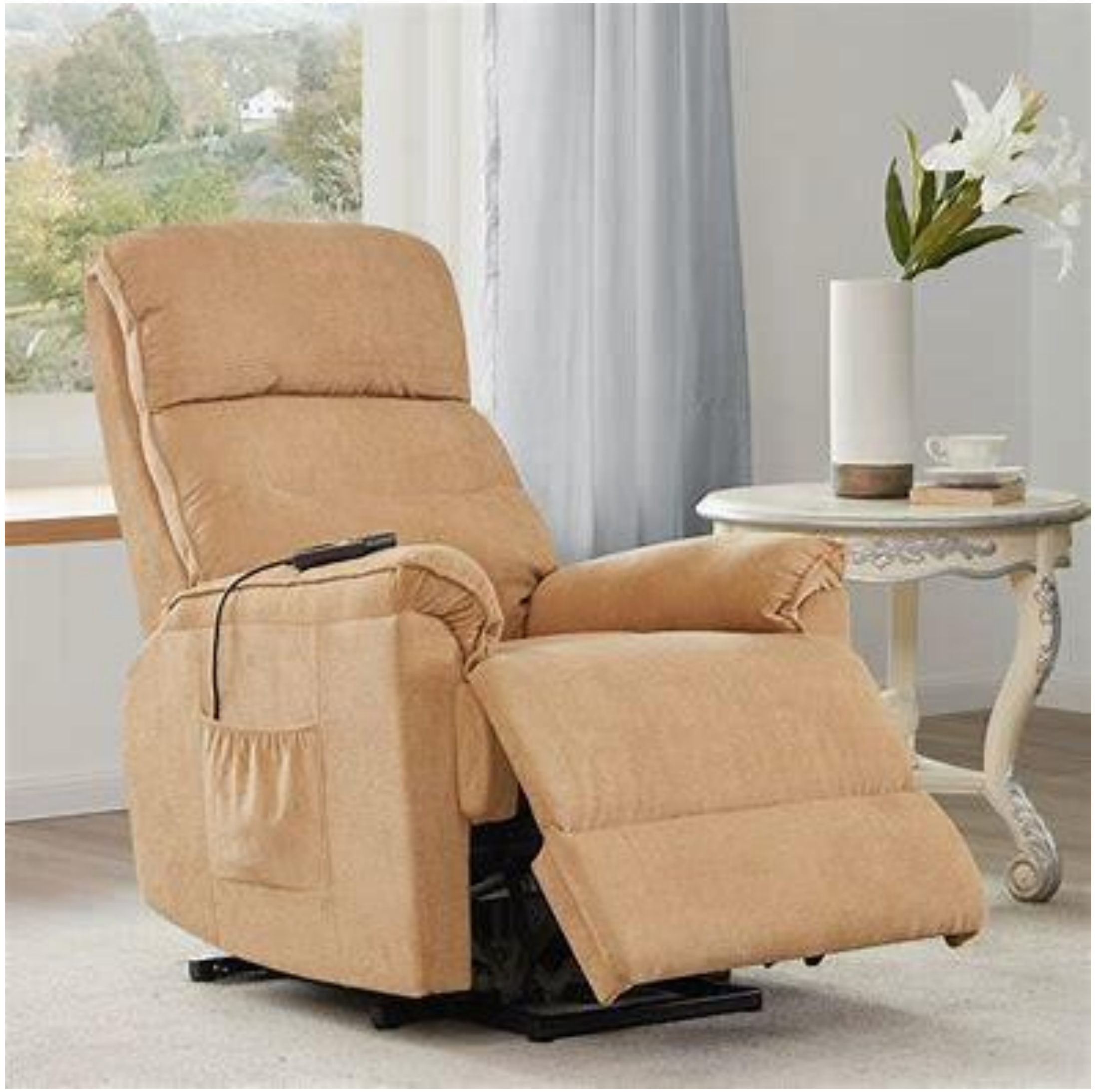
Chair 8



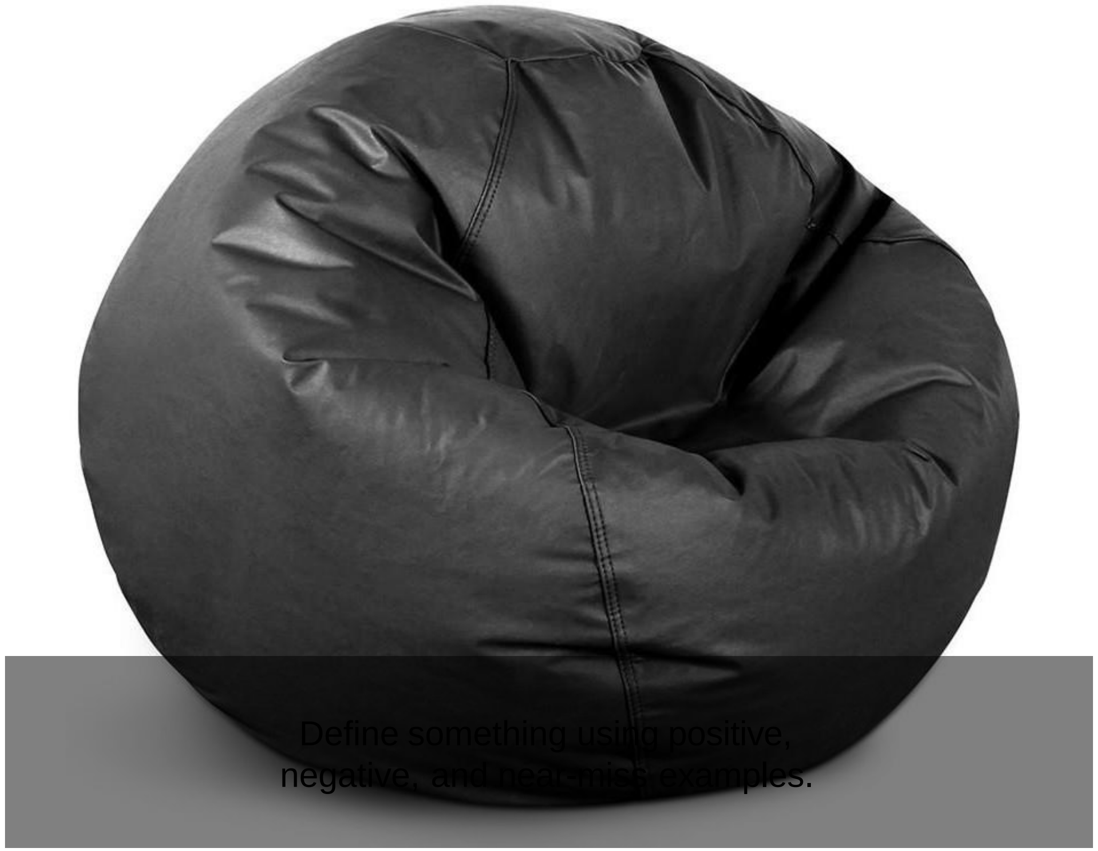
Chair 9



Chair ?10

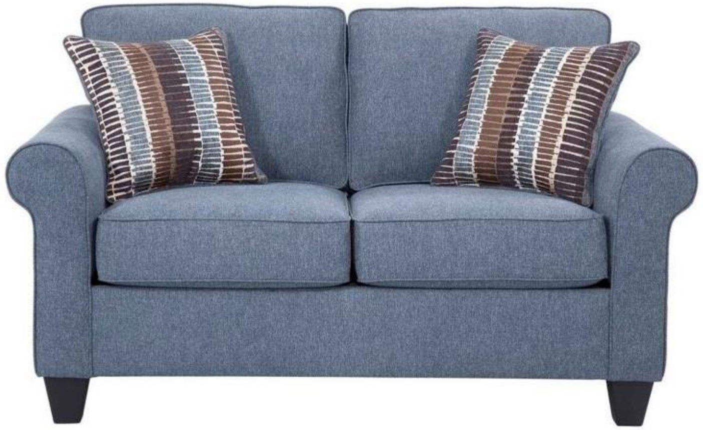


Chair ?11

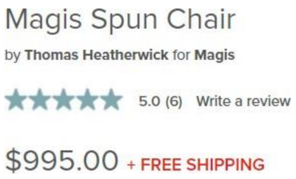


Define something using positive, negative, and near-miss examples.

Chair ?12



Chair ?13

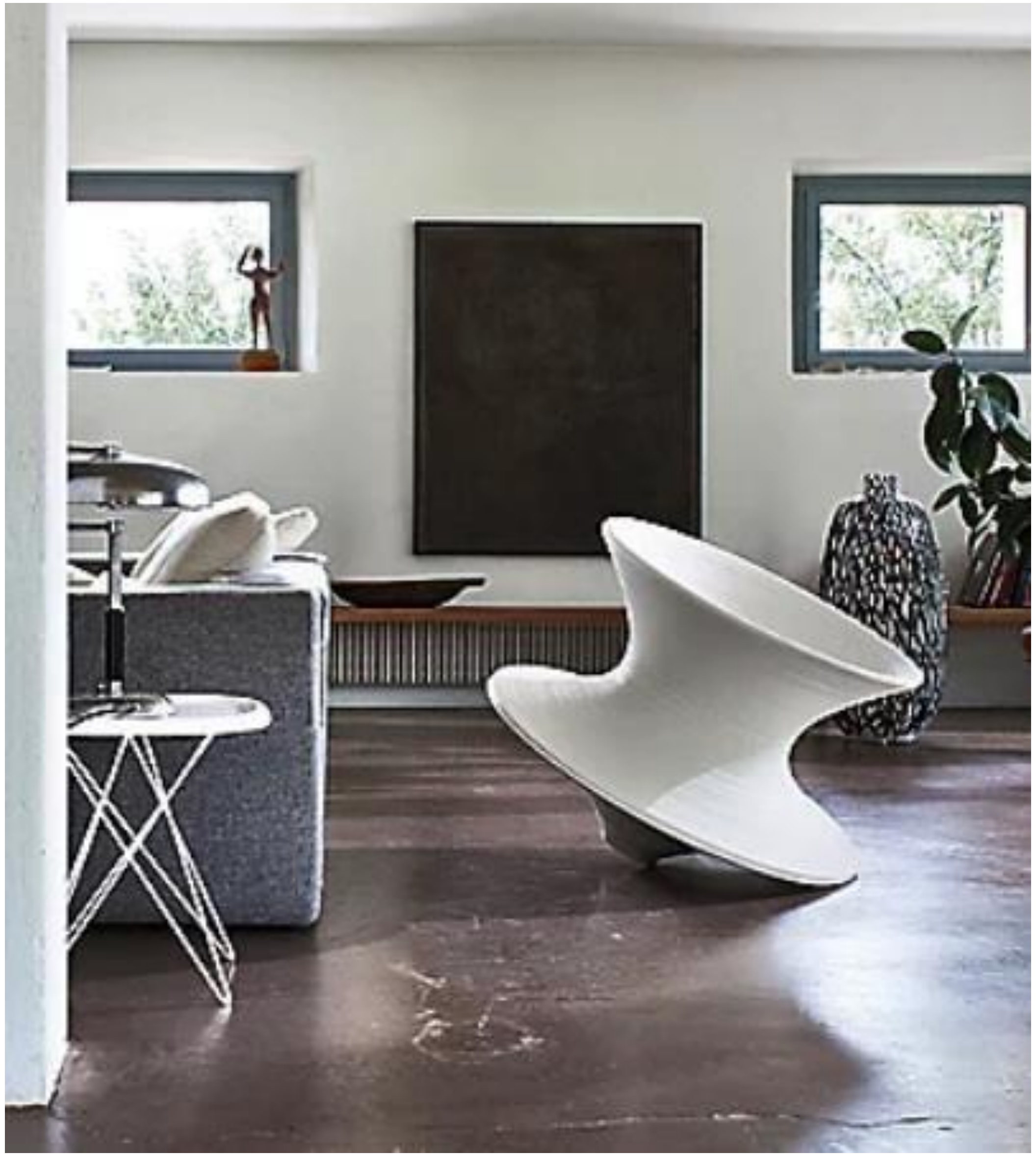


Magis Spun Chair

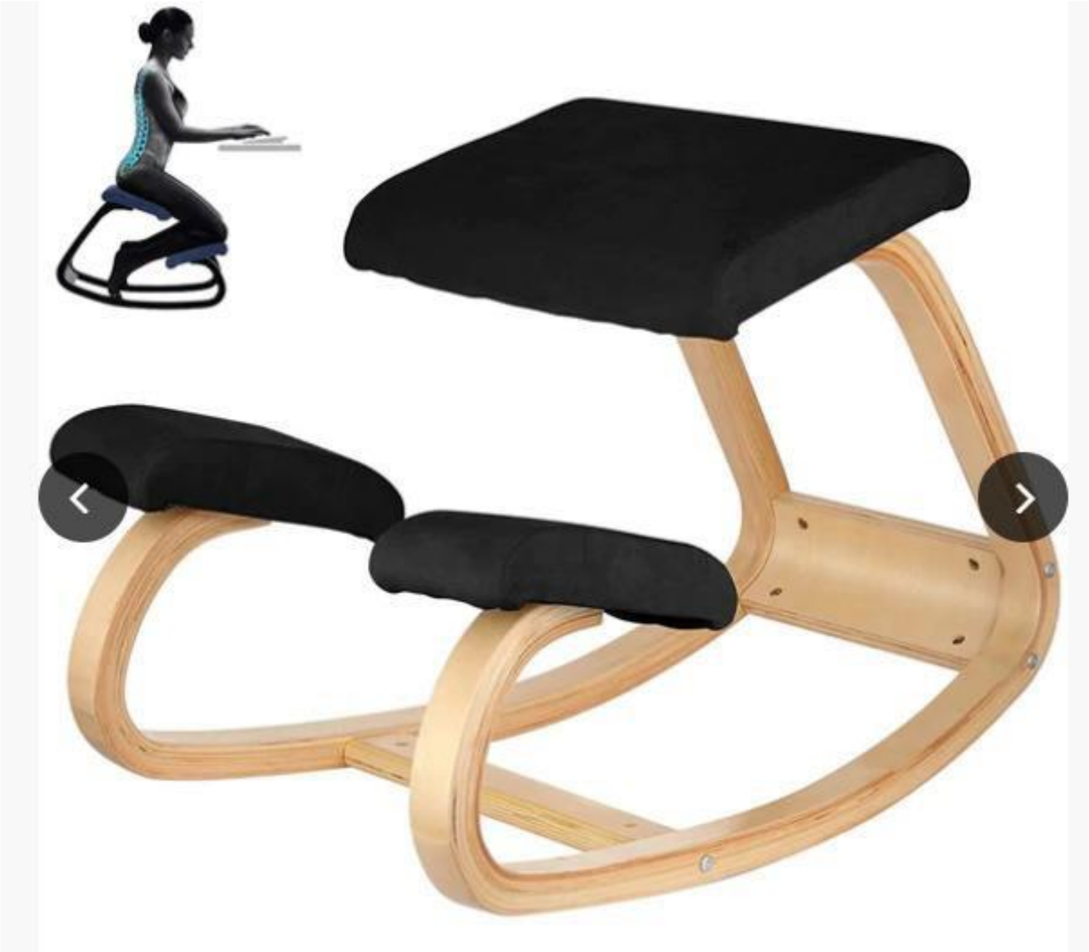
by **Thomas Heatherwick** for **Magis**

★★★★★ 5.0 (6) Write a review

$995.00 + FREE SHIPPING

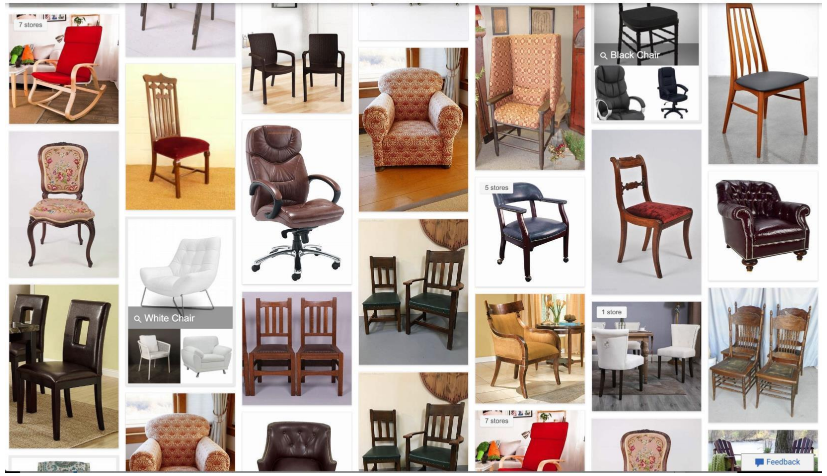


Chair ?14



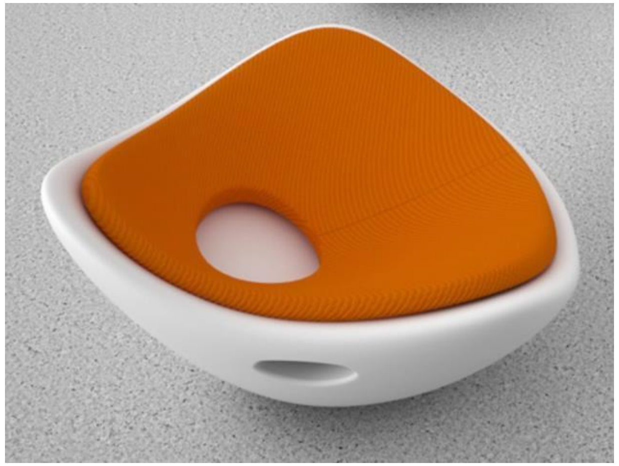
Ergonomic Kneeling Chair Posture Rocking Knee Stool For Home Office Meditation

How would you **model** a chair ? .. 3



- Sitting surface?
- Back?
- Kind of back?
- Legs?
- #Legs?
- Arms?
- Height adjustment?
- Back adjustment?
- Material? …

Introspection: Let’s analyze what we said/thought



Balance Chair on Behance

- The more precise we are in our description,
  the fewer chairs we describe
  and/or
  the longer the description

- Relaxing some ‘conditions’
  (e.g., chairs have 4 legs ➔ chairs have legs)
  is useful;
  relaxing it even more is better
  (e.g., chairs have legs ➔ chairs have some support) !

Task: How do we get a good-enough description
to include (most) chairs but exclude (most) non-chairs?

Think of models as compressed descriptions!

Optimization & Generalization

Tension between Optimization & Generalization

Optimization: adjusting model to get best perf on training data (‘learning’ from data)

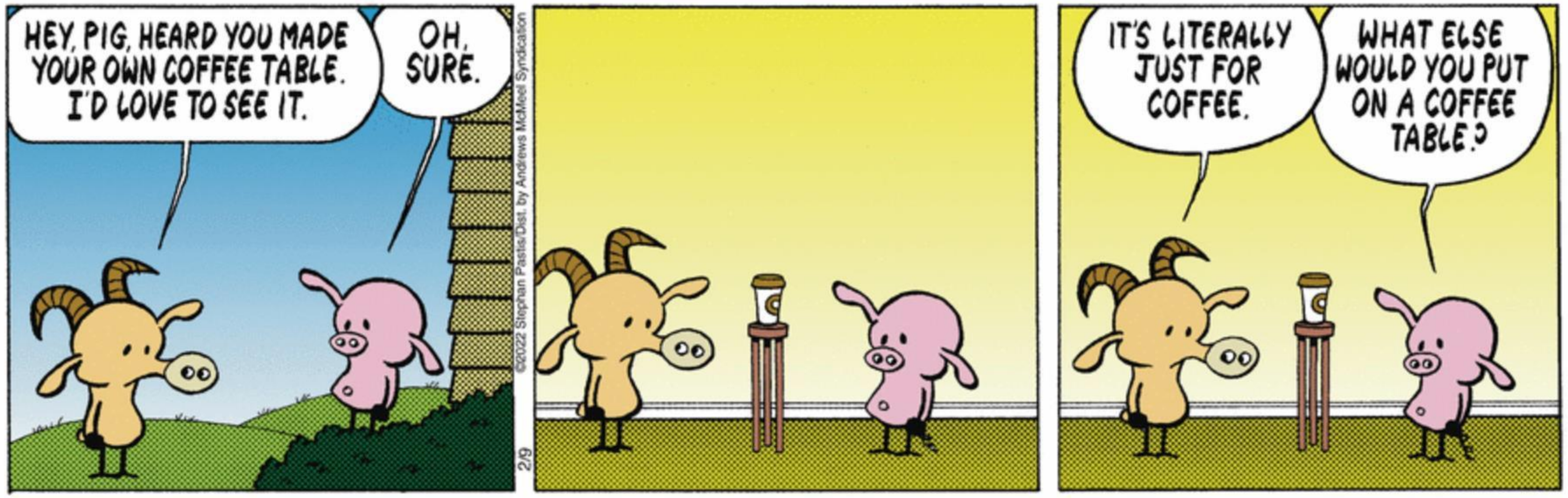
Generalization: how this model behaves on unseen data

Goal: Good generalization. But:

- Only model fitting is under our control
- If we fit too well, we get overfitting, and generalization suffers

From a recent newspaper…

Pearls Before Swine by Stephan Pastis for February 09, 2022


HEY, PIG, HEARD YOU MADE YOUR OWN COFFEE TABLE. I'D LOVE TO SEE IT.
OH, SURE.
©2022 Stephan Pastis/Dist. by Andrews McMeel Syndication
2/9
IT'S LITERALLY JUST FOR COFFEE.
WHAT ELSE WOULD YOU PUT ON A COFFEE TABLE?

$$\begin{aligned}
p(C_k, x_1, \ldots, x_n) &= p(x_1, \ldots, x_n, C_k) \\
&= p(x_1 \mid x_2, \ldots, x_n, C_k)\, p(x_2, \ldots, x_n, C_k) \\
&= p(x_1 \mid x_2, \ldots, x_n, C_k)\, p(x_2 \mid x_3, \ldots, x_n, C_k)\, p(x_3, \ldots, x_n, C_k) \\
&= \cdots \\
&= p(x_1 \mid x_2, \ldots, x_n, C_k)\, p(x_2 \mid x_3, \ldots, x_n, C_k) \cdots p(x_{n-1} \mid x_n, C_k)\, p(x_n \mid C_k)\, p(C_k)
\end{aligned}$$

$$p(x_i \mid x_{i+1}, \ldots, x_n, C_k) = p(x_i \mid C_k)$$

$$\begin{aligned}
p(C_k \mid x_1, \ldots, x_n) &\varpropto p(C_k, x_1, \ldots, x_n) \\
&= p(C_k)\, p(x_1 \mid C_k)\, p(x_2 \mid C_k)\, p(x_3 \mid C_k)\, \cdots \\
&= p(C_k) \prod_{i=1}^{n} p(x_i \mid C_k)\,,
\end{aligned}$$