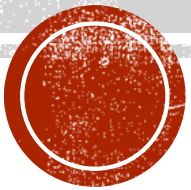
INSTANCE-BASED LEARNING AND NEAREST NEIGHBORS



CS 412 Introduction to Machine Learning

Prof. Zheleva

January 30, 2018

Reading Assignment: CIML: 3, ISL: 2.2.3

LAST LECTURE: LIMITS OF LEARNING

- Fundamental machine learning concepts
  - Bayes Optimal Classifier
  - What inductive bias is and what is its role in learning
  - What underfitting and overfitting means
  - How to estimate error on unseen examples
    - Why you should never touch your test data

BAYES OPTIMAL CLASSIFIER

- What if we had full access to the underlying data distribution D over (**x**,y) pairs?
- Bayes optimal classifier
  - For any input $\hat{x}$, we can compute the label $\hat{y}$

$$f^{(\mathrm{BO})}(\hat{x}) = \arg\max_{\hat{y} \in \mathcal{Y}} \mathcal{D}(\hat{x}, \hat{y})$$

Theorem 1 (Bayes Optimal Classifier). *The Bayes Optimal Classifier $f^{(BO)}$ achieves minimal zero/one error of any deterministic classifier.*

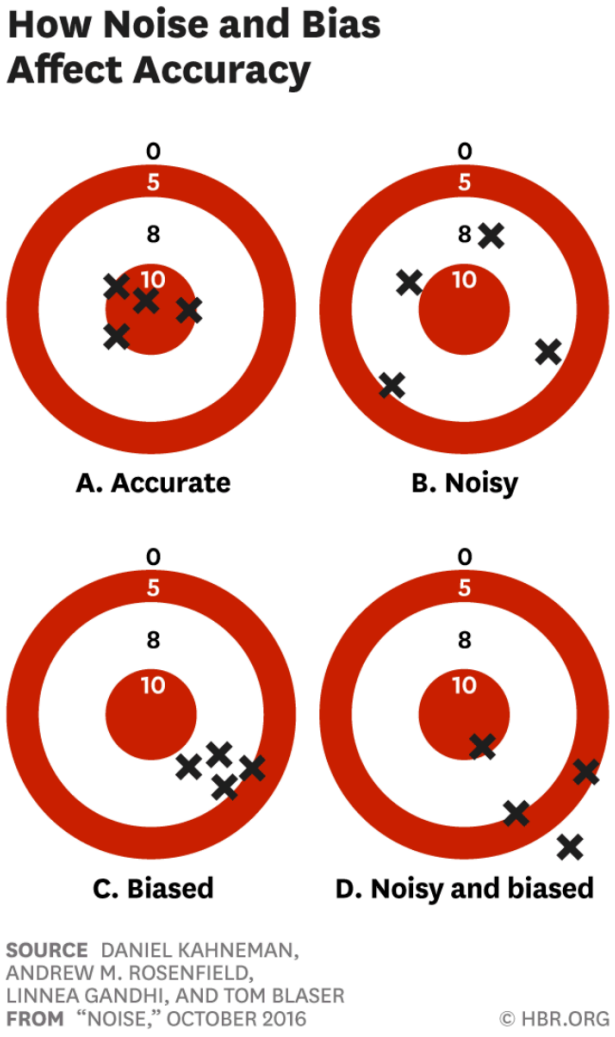
- Bayes error rate: error of Bayes optimal classifier
- Unfortunately, we never have access to D – so what do we do?

NOISE VS. BIAS IN TRAINING DATA

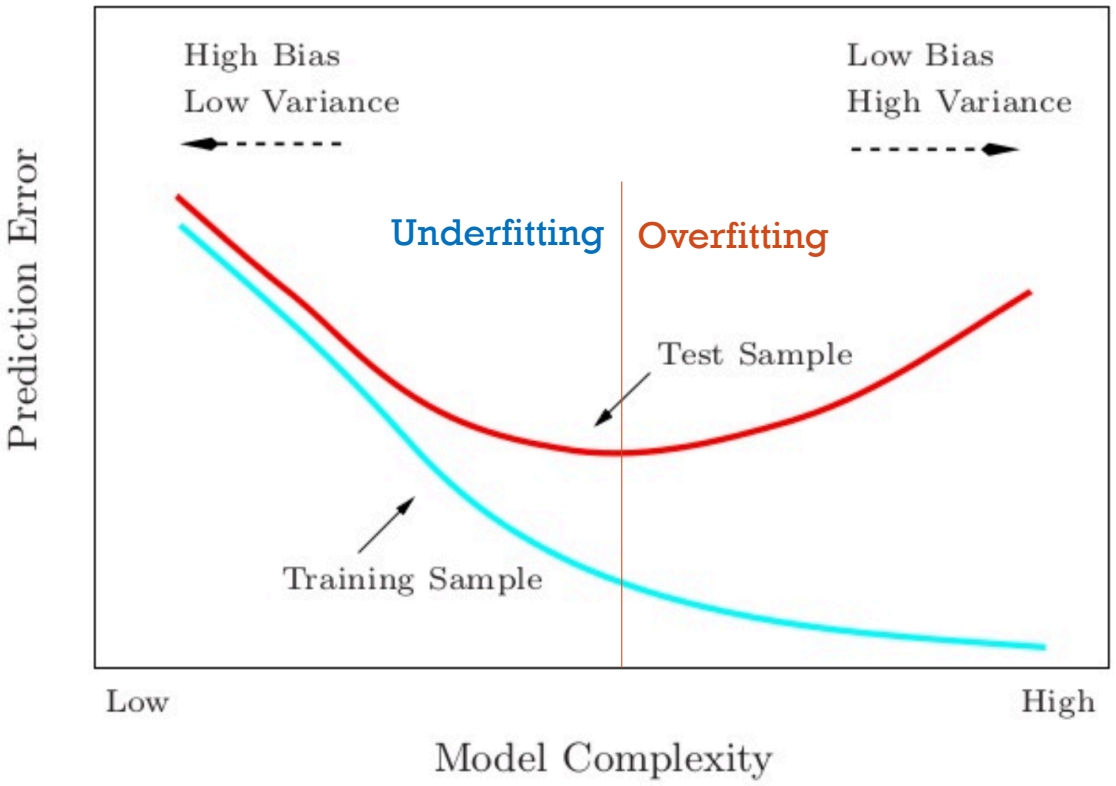
- Goal: estimate target in dart-throwing
- Training data: four sets of four examples

- Noise = variance

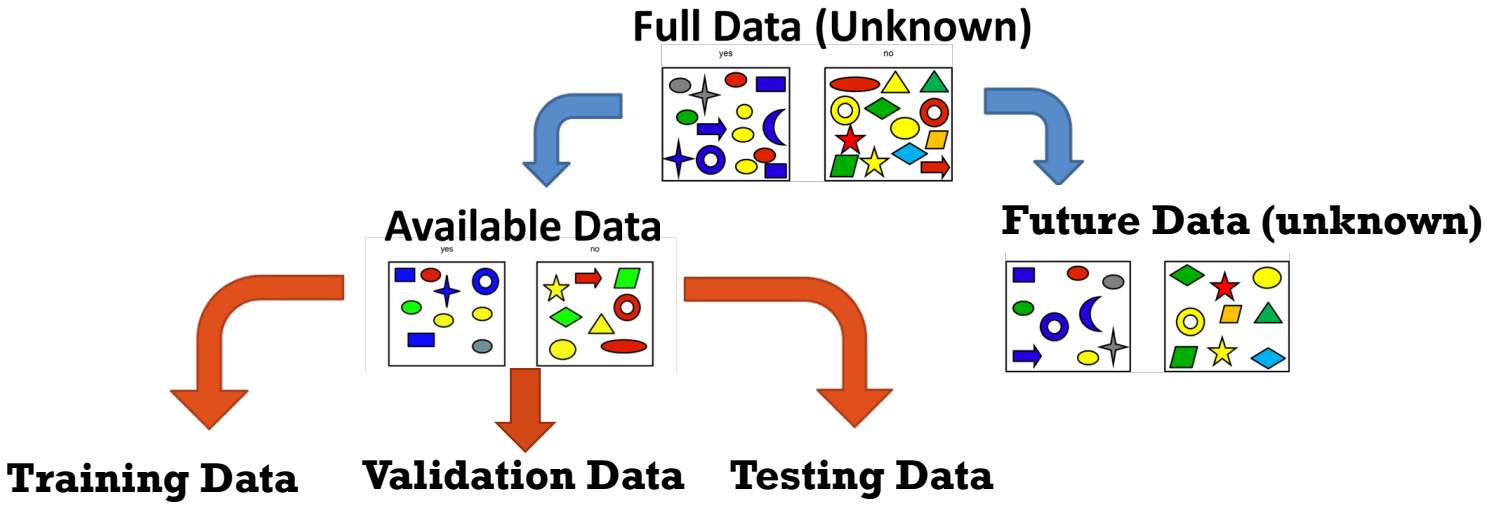
Example from https://hbr.org/2016/10/noise



UNDERFITTING VS. OVERFITTING



SUPERVISED LEARNING



We use testing data for evaluation

We don't look at testing data during training!

CARDINAL RULE OF MACHINE LEARNING

- Never ever touch your test data during learning!

- Can you check how noisy or biased your testing data is?
  - No!

- Can you tune your model parameters based on the testing data (e.g., figure out optimal tree depth)?
  - No!

- Can you visualize testing data to get insights on the differences between the training and testing data distribution?
  - No!

TODAY: INSTANCE-BASED LEARNING

- Practical concerns (finish from last time)
  - **Dealing with data: from raw data to well-defined examples**

- Nearest neighbor algorithms for classification
  - k-NN, weighted k-NN

- Fundamental machine learning concepts
  - Decision boundary

TRAINING A MODEL

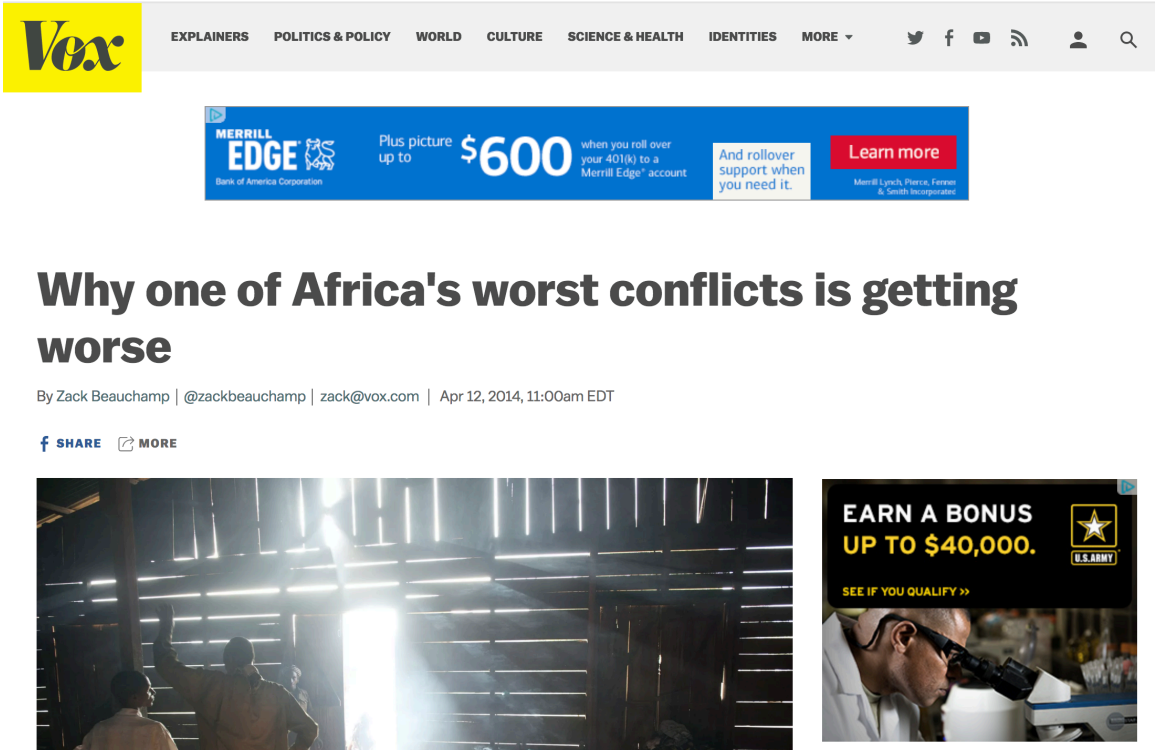
- Split your data into 70% training data, 10% development data and 20% test data.

- For each possible setting of your hyperparameters:
  - Train a model using that setting of hyperparameters on the training data
  - Compute this model's error rate on the development data.

- From the above collection of models, choose the one that achieved the lowest error rate on development data.

- Evaluate that model on the test data to estimate future test performance.

REAL-WORLD APPLICATIONS



1	real world goal	increase revenue
2	real world mechanism	better ad display
3	learning problem	classify click-through
4	data collection	interaction w/ current system
5	collected data	query, ad, click
6	data representation	bow^2, $\pm$ click
7	select model family	decision trees, depth 20
8	select training data	subset from april'16
9	train model & hyperparams	final decision tree
10	predict on test data	subset from may'16
11	evaluate error	zero/one loss for $\pm$ click
12	deploy!	(hope we achieve our goal)

WHAT ABOUT PREDICTING ARTICLE TOPIC?

- Real world goal
  - help readers find content
- Real world mechanism
  - tag articles with categories
- Simplified learning goal
  - is an article about "human rights" or not?
  - use binary classification
- Let's think about it in terms of decision trees
  - framework applies to other classifiers we will study

WHAT REAL DATA LOOKS LIKE

Why one of Africa's worst conflicts is getting worse

By Zack Beauchamp | @zackbeauchamp | zack@vox.com | Apr 12, 2014, 11:00am EDT

SHARE MORE

CAR civilians besieged by the anti-balaka militias take shelter under French military protection. | Miguel Medina/AFP/Getty Images

The conflict in the Central African Republic (CAR) is a huge international issue — thousands of people have been killed since it began in December 2012, and its worsening rapidly enough that the UN has just greenlit a

Input: x

Output: y

dsjVoxArticles.tsv

Download

	title	author	category
19	This slide is the best thing to come out of the Apple v. Samsung trial	Nilay Patel	Apple
20	40 states relaxed their drug laws in the past 5 years	German Lopez	Criminal Justice
21	The Dow topped 18,000 today. Learn why that matters.	Danielle Kurtzleben	Business & Finance
22	Statistical controls tell us how the gender pay gap works, not that it isn'	Matthew Yglesias	Politics & Policy
23	Don't panic yet, but early Obamacare enrollees might cost more	German Lopez	Health Care
24	The best evidence we have that Obamacare is working	Ezra Klein	Health Care
25	How does one doctor earn $21 million from Medicare?	Sarah Kliff	Health Care
26	Teens are shockingly great at using birth control	Sarah Kliff	Health Care
27	Some of you are lying about going to church	Brandon Ambrosino	Culture
28	How does one pill cost $1,000?	German Lopez	Health Care
29	HHS Secretary Sebelius to resign	Sarah Kliff	Health Care
30	Kathleen Sebelius is resigning because Obamacare has won	Ezra Klein	Health Care
31	Why one of Africa's worst conflicts is getting worse	Zack Beauchamp	Human Rights
32	Obama wants to fight gerrymandering once he leaves office. He should look t	Andrew Prokop	The Latest
33	5 disease outbreaks happening right now that vaccines could have prevented	German Lopez	Health Care
34	Do e-cigarettes kill you? Researchers donâ€™t know	German Lopez	Health Care
35	The $2.8 trillion question: Are health costs growing fast again?	Sarah Kliff	Health Care

Source: https://data.world/elenadata/vox-articles

dsjVoxArticles.tsv Download

	title	author	category
19	This slide is the best thing to come out of the Apple v. Samsung trial	Nilay Patel	Apple
20	40 states relaxed their drug laws in the past 5 years	German Lopez	Criminal Justice
21	The Dow topped 18,000 today. Learn why that matters.	Danielle Kurtzleben	Business & Finance
22	Statistical controls tell us how the gender pay gap works, not that it isn'	Matthew Yglesias	Politics & Policy
23	Don't panic yet, but early Obamacare enrollees might cost more	German Lopez	Health Care
24	The best evidence we have that Obamacare is working	Ezra Klein	Health Care
25	How does one doctor earn $21 million from Medicare?	Sarah Kliff	Health Care
26	Teens are shockingly great at using birth control	Sarah Kliff	Health Care
27	Some of you are lying about going to church	Brandon Ambrosino	Culture
28	How does one pill cost $1,000?	German Lopez	Health Care
29	HHS Secretary Sebelius to resign	Sarah Kliff	Health Care
30	Kathleen Sebelius is resigning because Obamacare has won	Ezra Klein	Health Care
31	Why one of Africa's worst conflicts is getting worse	Zack Beauchamp	Human Rights
32	Obama wants to fight gerrymandering once he leaves office. He should look t	Andrew Prokop	The Latest
33	5 disease outbreaks happening right now that vaccines could have prevented	German Lopez	Health Care
34	Do e-cigarettes kill you? Researchers donâ€™t know	German Lopez	Health Care
35	The $2.8 trillion question: Are health costs growing fast again?	Sarah Kliff	Health Care

1	real world goal	Help readers find content
2	real world mechanism	Tag articles with categories
3	learning problem	Is an article about "human rights"?
4	data collection	?
5	collected data	
6	data representation	
7	select model family	
8	select training data	
9	train model & hyperparams	
10	predict on test data	
11	evaluate error	
12	deploy!	

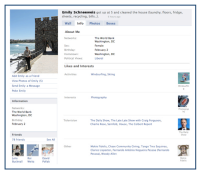
FROM INSTANCES TO FEATURE VECTORS X

- Text classification

Features: bag of words

aardvark	africa	beach	conflict	…	worse	zebra
0	1	0	1	…	1	0

- Voter classification



Features: personal attributes

Age	Gender	Religion	Hispanic	Golf	…
50	F	Catholic	No	No	

- Image classification

Features: image features

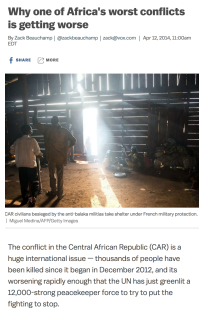
Eyes	Nose	Head	FaceColor	Hair
Round	Triangle	Round	Purple	Yes

NEXT: NEAREST NEIGHBORS

- Practical concerns (finish from last time)
  - Dealing with data: from raw data to well-defined examples
- **Nearest neighbor algorithms for classification**
  - k-NN
- Fundamental machine learning concepts
  - Decision boundary

GEOMETRIC VIEW OF DATA

- Each feature is a dimension
- The feature values of each data instance correspond to a point in n-dimensional space where n is the number of features



Features: bag of words

Feature values: presence of a word

aardvark	africa	beach	conflict	…	worse	zebra
0	1	0	1	…	1	0

Feature values: count of a word

aardvark	africa	beach	conflict	…	worse	zebra
0	5	0	2	…	1	0

INTUITION FOR NEAREST NEIGHBOR CLASSIFICATION

"This 'rule of nearest neighbor' has considerable elementary intuitive appeal and probably corresponds to practice in many situations. For example, it is possible that much medical diagnosis is influenced by the doctor's recollection of the subsequent history of an earlier patient whose symptoms resemble in some way those of the current patient."

(Fix and Hodges, 1952)

INTUITION FOR NEAREST NEIGHBOR CLASSIFICATION

- Simple idea
  - Store all training examples
  - Classify new examples based on most similar examples

K NEAREST NEIGHBOR

Algorithm 3 KNN-PREDICT($\mathbf{D}$, K, $\hat{x}$)

1: $S \leftarrow [\,]$
2: **for** $n = 1$ **to** N **do**
3: $\quad S \leftarrow S \oplus \langle \mathrm{d}(x_n, \hat{x}), n \rangle$ // store distance to training example n
4: **end for**
5: $S \leftarrow \text{SORT}(S)$ // put lowest-distance objects first
6: $\hat{y} \leftarrow 0$
7: **for** $k = 1$ **to** K **do**
8: $\quad \langle dist, n \rangle \leftarrow S_k$ // n this is the kth closest data point
9: $\quad \hat{y} \leftarrow \hat{y} + y_n$ // vote according to the label for the nth training point
10: **end for**
11: **return** SIGN($\hat{y}$) // return $+1$ if $\hat{y} > 0$ and -1 if $\hat{y} < 0$

D: training data
K: number of neighbors used for classification
$\hat{x}$: test instance with unknown class in {-1,+1}

TWO TYPES OF LEARNING

Eager learning

Example: decision trees

- Learn/train:
  - Induce an abstract model from data
- Test/predict/classify:
  - Apply learned model to new data
- Properties
  - Retains a model and its parameters
    - Requires time to train
  - Classification is typically much faster

Instance-based learning

Example: k-NN

- Learn:
  - Just store data in memory
- Test/predict/classify:
  - Compare new data to **all** stored data
- Properties
  - Retains all information seen in training data
    - Doesn't need time to train
  - Classification can be very slow

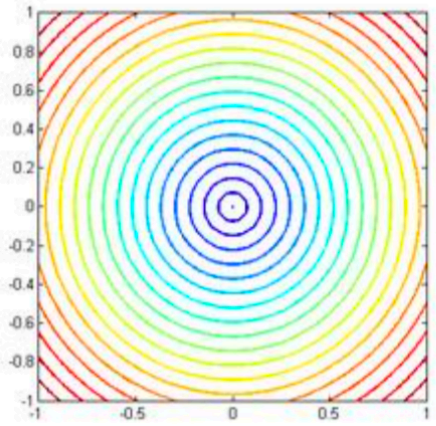
COMPONENTS OF A K-NN CLASSIFIER

- **Distance metric**
  - How do we measure distance between instances
  - Determines the layout of the example space
- The k hyperparameter
  - How large a neighborhood should we consider?
  - Determines the complexity of the hypothesis space

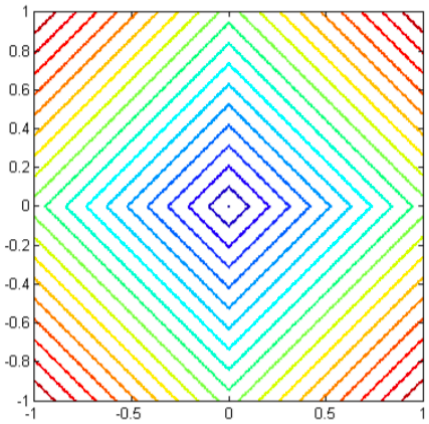
DISTANCE METRIC

- L2 norm or Euclidean distance: $||\boldsymbol{a} - \boldsymbol{b}||_2 = \sqrt{\sum_i (a_i - b_i)^2}$ where i is the dimension
- L1 norm or Manhattan distance: $||\boldsymbol{a} - \boldsymbol{b}||_1 = \sum_i |ai - bi|$
- Max norm or Chebyshev distance: $||\boldsymbol{a} - \boldsymbol{b}||_\infty = maxi(|a_i - bi|)$
- Different distances yield different neighborhoods

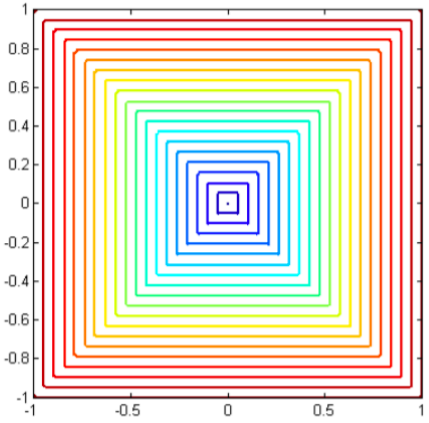
L2 distance
(= Euclidean distance)

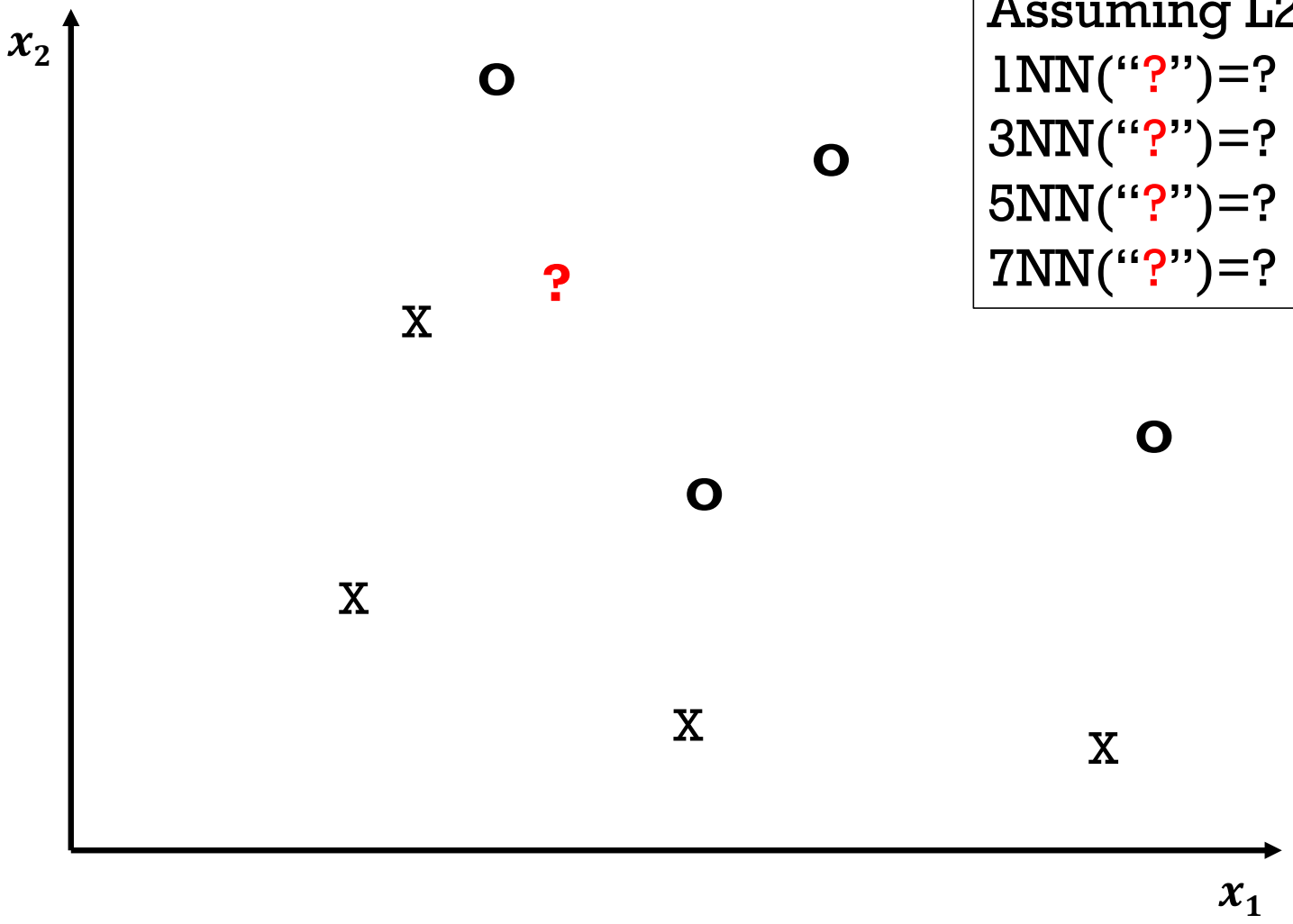


L1 distance



Max norm





Assuming L2 norm:
1NN("?")=?
3NN("?")=?
5NN("?")=?
7NN("?")=?

CATEGORICAL FEATURES?

Eyes	Nose	Head	FaceColor	Hair
Round	Triangle	Round	Purple	Yes

Nose ϵ {Round,Triangle,Square}

$\mathbf{a}_j, \mathbf{b}_j \epsilon$ {dog, cat, tree, …}

Indicator function: creates new variables (1-hot encoding)

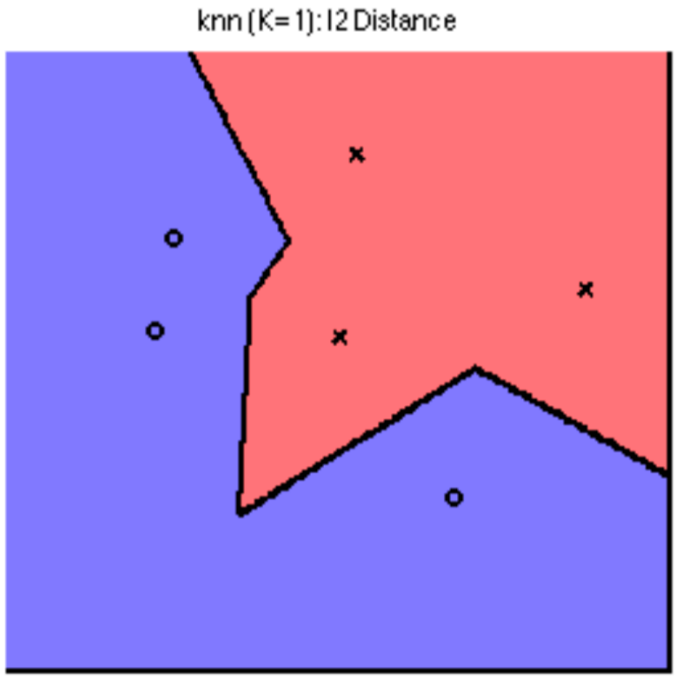
Eyes-Round	Nose-Round	Nose-Triangle	Nose-Square…
1	0	1	0

DECISION BOUNDARY OF A CLASSIFIER

- The line that separates positive and negative regions in the feature space

- Why is it useful?
  - It helps us visualize how examples will be classified for the entire feature space
  - It helps us visualize the complexity of the learned model

- Decision boundary for 1NN →

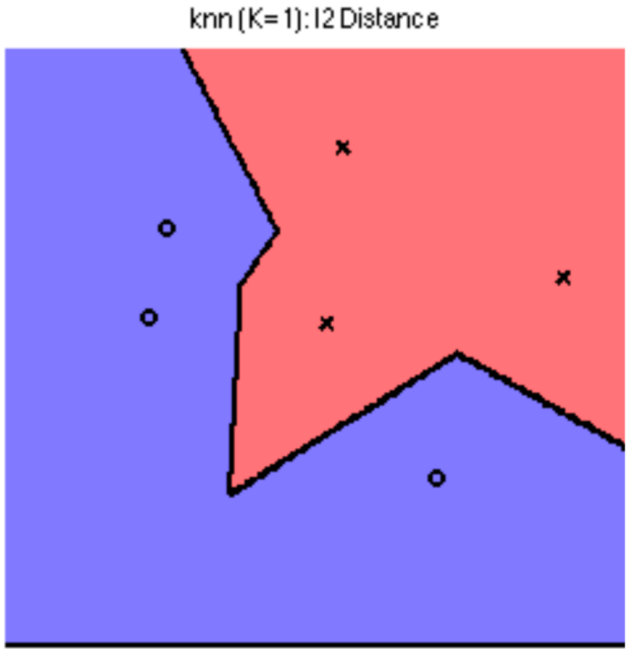


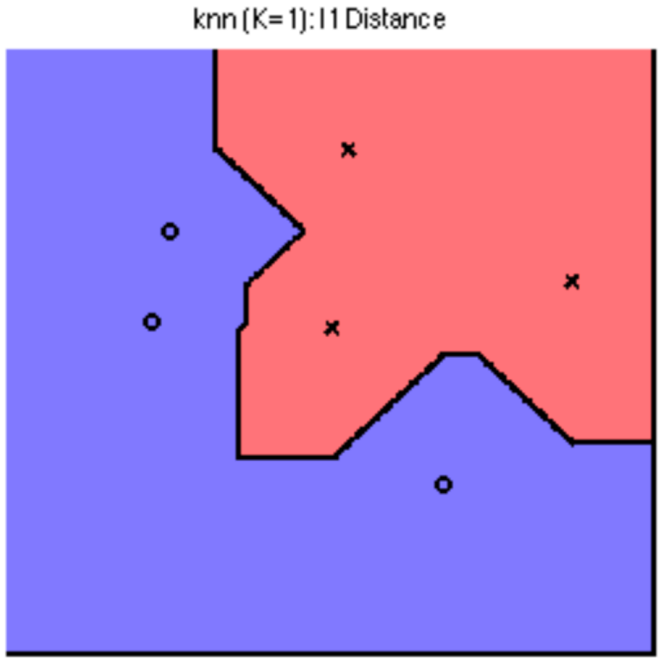
FINDING DECISION BOUNDARIES FOR 1NN

- Can be done in O(nlogn) using Voronoi diagrams (dual of Delaunay triangulation)

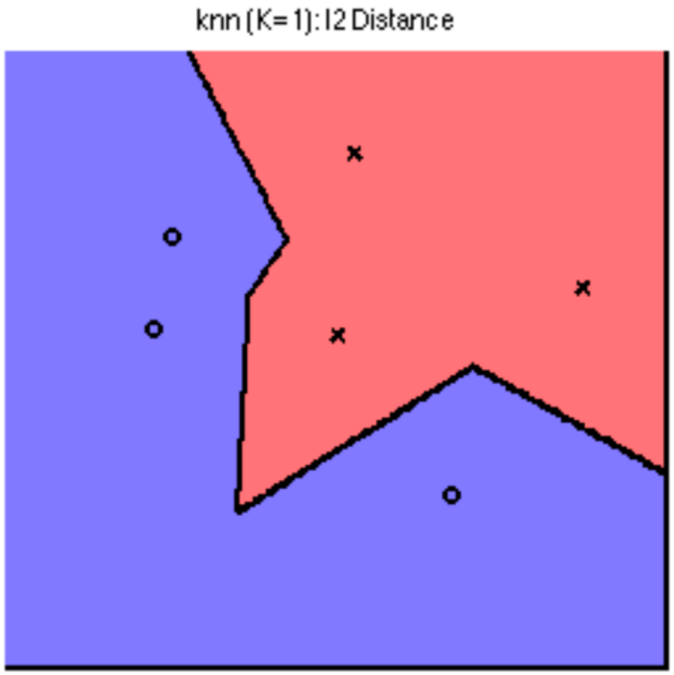
Interactive Voronoi diagram: http://alexbeutel.com/webgl/voronoi.html

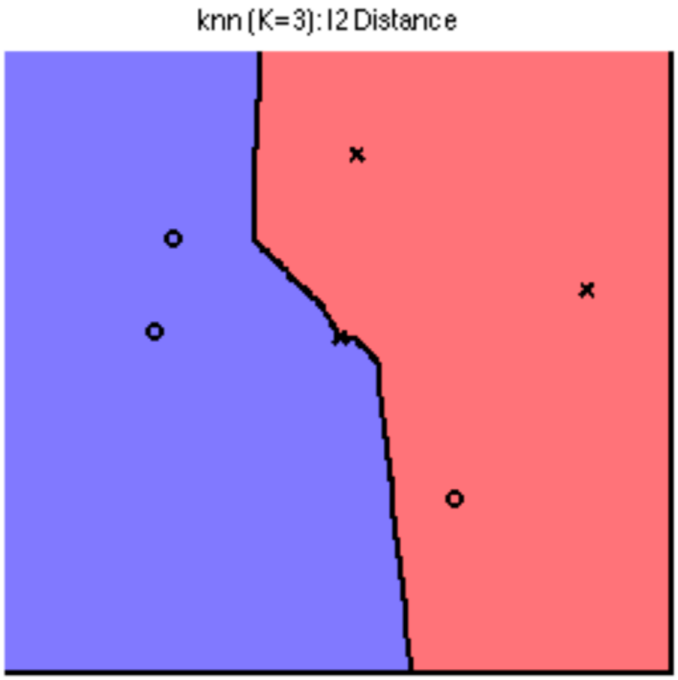
DECISION BOUNDARIES DEPEND ON THE DISTANCE FUNCTION





DECISION BOUNDARIES CHANGE WITH K

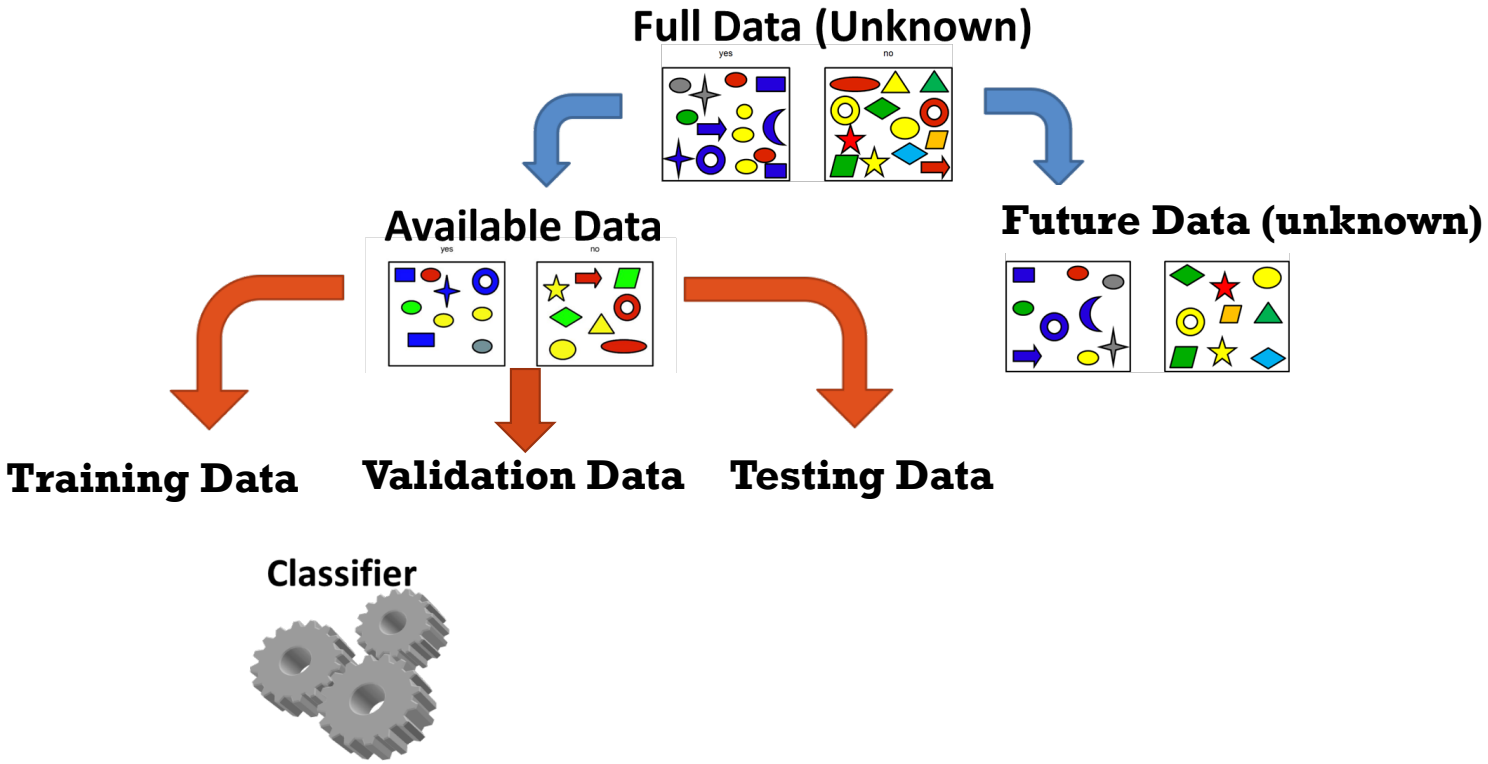




THE K HYPERPARAMETER

- Tunes the complexity of the hypothesis space
  - If k = 1, every training example has its own neighborhood
  - If k = N, the entire feature space is one neighborhood!
- Higher k yields smoother decision boundaries
- How would you set k in practice?

SUPERVISED LEARNING



INDUCTIVE BIAS OF KNN

- Nearby instances should have the same label
- All features are equally important (very different from decision trees!)
- Complexity is tuned by the k parameter

VARIANT ON KNN: WEIGHTED VOTING

- By default, all neighbors have equal weight
- Instance-weighted voting: weight neighbors by distance
  - For example, in inverse distance-weighted voting, the weight of each neighbor vote is

$$w_i = f(dist(\hat{x}, x_i)) = \begin{cases} \infty & if\ dist(\hat{x}, x_i) = 0 \\ \frac{1}{dist(\hat{x}, x_i)} & otherwise \end{cases}$$

WEIGHTED KNN

Algorithm 3 KNN-PREDICT($\mathbf{D}$, K, $\hat{x}$)

1: $S \leftarrow [\,]$
2: **for** $n = 1$ **to** N **do**
3: $\quad S \leftarrow S \oplus \langle \mathrm{d}(x_n, \hat{x}), n \rangle$ // store distance to training example n
4: **end for**
5: $S \leftarrow$ SORT(S) // put lowest-distance objects first
6: $\hat{y} \leftarrow 0$
7: **for** $k = 1$ **to** K **do**
8: $\quad \langle dist, n \rangle \leftarrow S_k$ // n this is the kth closest data point
9: $\quad \hat{y} \leftarrow \hat{y} + f(dist) y_n$ // vote according to the label for the nth training point
10: **end for**
11: **return** SIGN($\hat{y}$) // return $+1$ if $\hat{y} > 0$ and -1 if $\hat{y} < 0$

D: training data
K: number of neighbors used for classification
$\hat{x}$: test instance with unknown class in {-1,+1}

CURSE OF DIMENSIONALITY

- Challenges of working with high dimensional spaces
  - Hard to visualize
  - Computational cost
  - Many of our intuitions about 2D or 3D spaces don't hold
    - High dimensional hyperspheres "look more like porcupines than balls"
    - Distances between two random points in high dimensions are approximately the same

EXERCISE: DECISION TREES VS. KNN

Properties of classification problem	Can Decision Trees handle them?	Can K-NN handle them?
Binary features		
Numeric features		
Categorical features		
Robust to noisy training examples		
Fast classification is crucial		
Many irrelevant features		
Relevant features have very different scale		

SUMMARY

- From real-world problem to data to model to solution
- Using geometry for classification
- kNN: Very simple non-linear classification technique
  - Works well with large amounts of data
  - Related to very sophisticated machine learning technique: kernel methods!
- Decision boundaries

ANNOUNCEMENTS

- HW 1 is due tonight

ACKNOWLEDGEMENTS

- These slides use materials by Brian Ziebart, Marine Carpuat