وزارت علوم تحقیقات و فناوری
دانشگاه علم و فرهنگ

Machine Learning

Lecture 4.
Supervised learning
Naïve Bayes

Alireza Rezvanian
Fall 2022

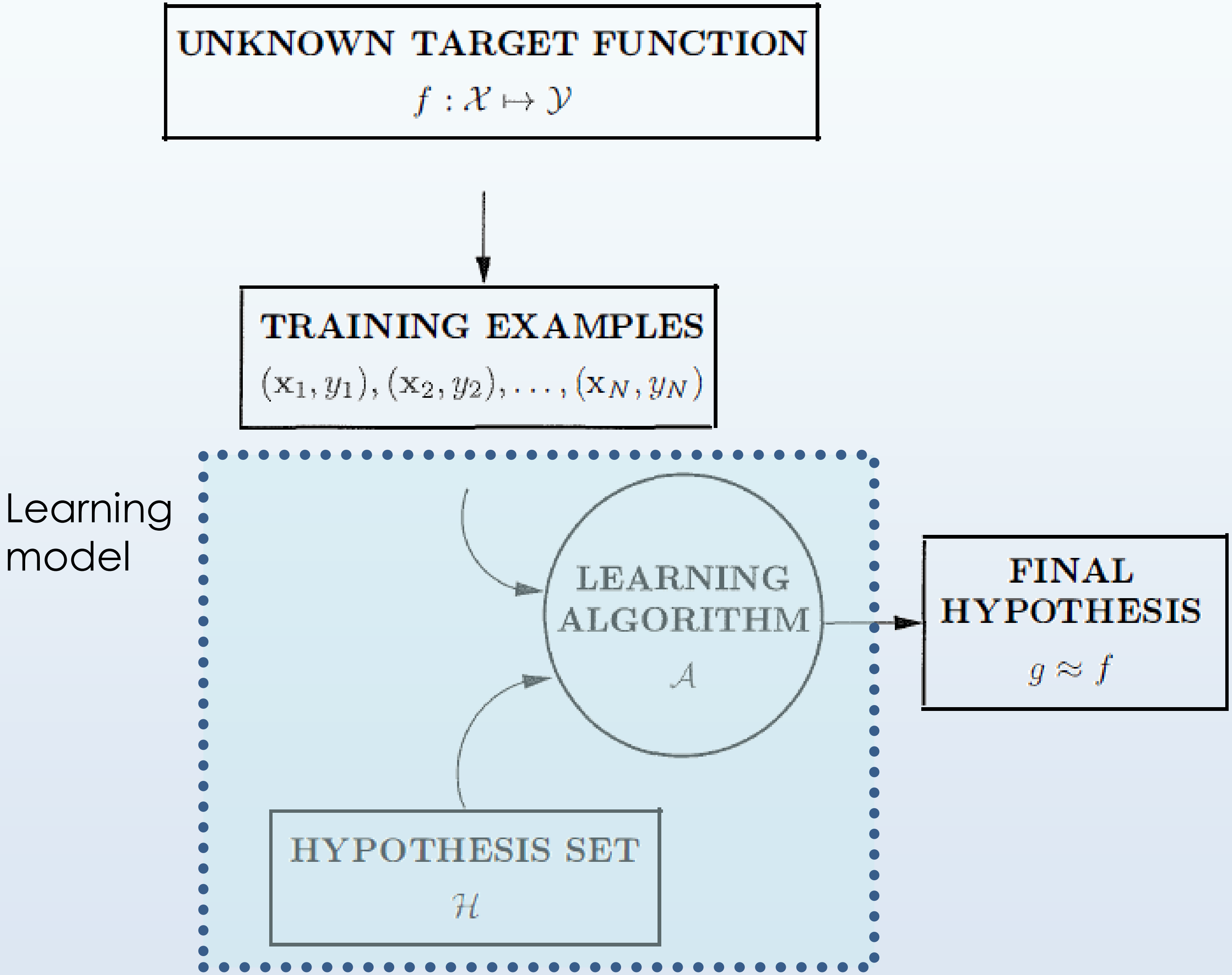
Amirkabir University of Technology (Tehran Polytechnic)

Last update: Oct. 22, 2021

Outline

- Probability and Inference
- Bayes' Rule
- MAP & MLE
- Naive Bayes Classifier
- m-estimate
- Text classification

Supervised Learning



Statistical Estimation View

- Probabilities to rescue:
  - x and *y* are *random variables*
  - $D = (x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N) \quad \sim P(X,Y)$
- IID: **I**ndependent **I**dentically **D**istributed
  - Both training & testing data sampled IID from P(X,Y)
  - Learn on training set
  - Have some hope of *generalizing* to test set

Interpreting Probabilities

- What does P(A) mean?

- Frequentist View
  - limit $N \rightarrow \infty$ #(A is true)/N
  - limiting frequency of a repeating non-deterministic event

- Bayesian View
  - P(A) is your "belief" about A

- Market Design View
  - P(A) tells you how much you would bet

Concepts

- Likelihood: $P(D|h)$
  - How much does a certain hypothesis explain the data?

- Prior: $P(h)$
  - What do you believe before seeing any data?

- Posterior: $P(h|D)$
  - What do we believe after seeing the data?

Classification

- **Learn**: h:**X** $\mapsto$ Y
  - **X** – features
  - Y – target classes

- Suppose you know P(Y | **X**) exactly, how should you classify?
  - Bayes classifier:

- **Why?**

Probability and Inference

- Result of tossing a coin is $\in$ {Heads,Tails}
- Random var $X \in \{1,0\}$

  Bernoulli: $P\{X=1\} = p_o^X (1 - p_o)^{(1-X)}$
- Sample: $\boldsymbol{X} = \{x^t\}_{t=1}^N$

  Estimation: $p_o = \#\{\text{Heads}\}/\#\{\text{Tosses}\} = \sum_t x^t / N$
- Prediction of next toss:

  Heads if $p_o > ½$, Tails otherwise

Classification

- **Credit scoring**: Inputs are income and savings. Output is *low-risk* vs. *high-risk*
- **Input**: $\mathbf{x} = [x_1, x_2]^T$,Output: $C \in \{0,1\}$
- **Prediction**:

$$\text{choose} \begin{cases} C = 1 \text{ if } P(C=1 \mid x_1, x_2) > 0.5 \\ C = 0 \text{ otherwise} \end{cases}$$

or

$$\text{choose} \begin{cases} C = 1 \text{ if } P(C=1 \mid x_1, x_2) > P(C=0 \mid x_1, x_2) \\ C = 0 \text{ otherwise} \end{cases}$$

Bayes Theorem

posterior *likelihood* *prior*

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

evidence

P(h): Prior probability of hypothesis h (Prior)

P(D): Prior probability of training data D (Evidence)

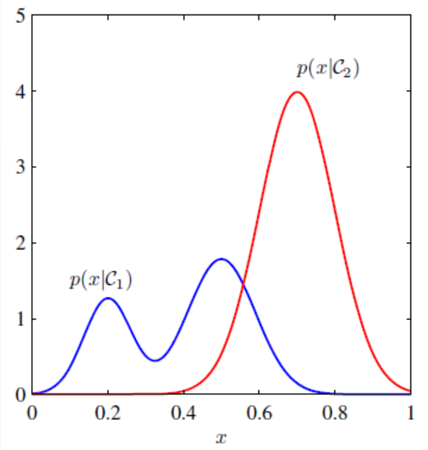
P(h | D): Probability of h given D (Posterior)

P(D | h): Probability of D given h (Likelihood)

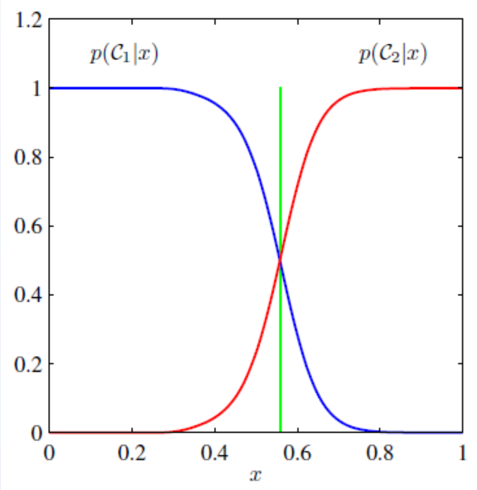
Generative vs. Discriminative

- Using Bayes rule, optimal classifier

$$h^*(x) = \underset{c}{\operatorname{argmax}}\{log p(x|y=c) + log p(y=c)\}$$



- Generative Approach
  - Estimate **p(x|y)** and **p(y)**
  - Use Bayes Rule to predict **y**



- Discriminative Approach
  - Estimate **p(y|x)** directly OR
  - Learn "discriminant" function **h(x)**

Choosing hypothesis

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

- Generally we want the most probable hypothesis given the training data
  - Maximum A Posteriori (MAP) hypothesis $\mathbf{h_{MAP}}$

$$\begin{aligned} h_{MAP} &\equiv \arg\max_{h \in H} P(h \mid D) \\ &= \arg\max_{h \in H} \frac{P(D \mid h)P(h)}{\cancel{P(D)}} \\ &= \arg\max_{h \in H} P(D \mid h)P(h) \end{aligned}$$

Evidence is a Constant independent of ***h***

- If assume $P(h_i) = P(h_j)$ then can further simplify, and choose
  - Maximum likelihood (ML) hypothesis

$$\begin{aligned} h_{ML} &\equiv \arg\max_{h \in H} P(h \mid D) \\ &= \arg\max_{h \in H} \frac{P(D \mid h)\cancel{P(h)}}{\cancel{P(D)}} \\ &= \arg\max_{h \in H} P(D \mid h) \end{aligned}$$

No prior hypothesis
OR

equally probable a priori ($P(h_i) = P(h_j)$)

Example: a medical diagnosis problem

- There are two alternative hypotheses:
  - h(1) that the patient has a particular form of cancer. (**+**)
  - h(2) that the patient does no (**-**)
- Prior knowledge indicates that over the entire population only **0.008** have this disease.
  - P(cancer) = 0.008 P(~cancer) = 0.992
- The test returns a correct **positive result** in only **98%** of the actual patent
  - P(+| cancer) = 0.98 P(-| cancer) = 0.02
- The test returns a correct **negative result** in only **97%** of the not patent
  - P(-| ~cancer) = 0.97 P(+|~cancer) = 0.03

Medical diagnosis problem

- Suppose we now observe a new patient for whom the lab test returns a **positive result**. Should we diagnose the patient as having cancer or not? **P(cancer|+)=?**
  - The maximum a posteriori (MAP) hypothesis

    $P(+|\ cancer) \times P(cancer) = (0.98) \times (0.008) = 0.0078$

    $P(+|\ \sim cancer) \times P(\sim cancer) = (0.03) \times (0.992) = 0.0298$

- Thus, $h_{MAP} = \sim cancer$

  Note: The exact posterior probabilities by normalizing

$$P(cancer|+) = \frac{0.0078}{0.0078 + 0.0298} = 0.21$$

Brute-force MAP Learning

1) For each hypothesis *h* in *H*, calculate the posterior probability

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

2) Output the hypothesis $\boldsymbol{h_{MAP}}$ with the highest posterior probability

$$h_{MAP} = \underset{h \in H}{\operatorname{argmax}}\, P(h|D)$$

Naive Bayes Classifier

- Target function: f: X → V
- tuple of attribute values **<a_1, a_2,…, a_n>**
- ***Goal:*** predict the target value, or classification, for this new instance.
- The Bayesian approach is to assign the most probable target value, V_{MAP}, given the attribute values <a_1, a_2,…, a_n> that describe the instance

$$V_{MAP} = \underset{v_j \in V}{\operatorname{argmax}} P(v_j | a_1, a_2, \ldots a_n)$$

$$V_{MAP} = \underset{v_j \in V}{\operatorname{argmax}} \frac{P(a_1, a_2, \ldots a_n | v_j) P(v_j)}{P(a_1, a_2, \ldots a_n)}$$

$$= \underset{v_j \in V}{\operatorname{argmax}} P(a_1, a_2, \ldots a_n | v_j) P(v_j)$$

Naive Bayes Classifier

- $V_{MAP} = \underset{v_j \in V}{\operatorname{argmax}} P(a_1, a_2, \dots a_n | v_j) P(v_j)$
  - ***P(v$_j$)*** computes by counting the frequency with which each target value ***v$_j$*** occurs in the training data
  - Estimating $\boldsymbol{P(a_1, a_2, \dots a_n | v_j)}$ is not feasible
- The **naive** Bayes classifier is based on the simplifying **assumption** that the attribute values are conditionally independent given the target value

$$P(a_1, a_2, \dots a_n | v_j) = P(a_1 | v_j) \times P(a_2 | v_j) \dots P(a_n | v_j) = \prod_i^n P(a_i | v_j)$$

$$\boldsymbol{V_{NB} = \underset{v_j \in V}{argmax}\, P(v_j) \prod_i^n P(a_i | v_j)}$$

- If assumption holds, NB is optimal classifier!

Naive Bayes Algorithm

Naïve_Bayes_Learn(*example*)

For each target value ***v***

For each target value $\boldsymbol{v_j}$

$\widehat{\boldsymbol{P}}(v_j) \leftarrow estimate\ P(v_j)$

For each attribute value $\boldsymbol{a_i}$ of each attribute a

$\widehat{\boldsymbol{P}}(a_i|v_j) \leftarrow estimate\ P(a_i|v_j)$

Classify_New_Instance(x)

$$\boldsymbol{V_{NB} = \underset{v_j \in V}{argmax}\left(\widehat{P}(v_j)\prod_{i}^{n}\widehat{P}(a_i|v_j)\right)}$$

Example: car theft problem

Example No.	Color	Type	Origin	Stolen?
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes

classify a **Red Domestic SUV** is getting stolen or not?

Example: car theft problem

- P(Stolen=Yes) = 5/10
- P(Stolen=No) = 5/10

Prior

- P(Color=Red | Stolen=Yes)=3/5
- P(Color=Red | Stolen=No)=2/5
- P(Type= SUV | Stolen=Yes)=1/5
- P(Type= SUV | Stolen=No)=3/5
- P(Origin=Domestic | Stolen=Yes)=2/5
- P(Origin=Domestic | Stolen=No)=3/5

Likelihood

- P(Stolen=Yes | Color=Red,Type=SUV,Origin=Domestic)=

  P(Color=Red |Stolen=Yes)×P(Type=SUV | Stolen=Yes) ×P(Origin =Domestic |Stolen=Yes)×P(Stolen=Yes)=?

  3/5 × 1/5 × 2/5 × 1/2 = 0.024

  $$\frac{0.024}{0.024+0.072} = 0.25$$

- P(Stolen=No | Color=Red,Type=SUV,Origin=Domestic)=

  P(Color=Red |Stolen=No)×P(Type=SUV | Stolen=No) ×P(Origin =Domestic |Stolen=No)×P(Stolen=No)=?

  2/5 × 3/5 × 3/5 × 1/2 = 0.072

Answer: the car is not stolen!

$$\frac{0.072}{0.024+0.072} = 0.75$$

Subtleties of NB classifier 1 – Violating the NB assumption

- Usually, features are not conditionally independent:

$$P(X_1 ... X_n|Y) \neq \prod_i P(X_i|Y)$$

- Probabilities P(Y | **X**) often biased towards 0 or 1

- Nonetheless, NB is a very popular classifier
  - NB often performs well, even when assumption is violated
  - [Domingos & Pazzani '96] discuss some conditions for good performance

Subtleties of NB classifier 2 – Insufficient training data

- What if you never see a training instance where X_1=a when Y=c?
  - e.g., Y={NonSpamEmail}, X_1={'Nigeria'}
  - $P(X_1=a \mid Y=c) = 0$
- Thus, no matter what the values $X_2,\ldots,X_d$ take:
  - $P(Y=c \mid X_1=a, X_2,\ldots,X_d) = 0$
- What now???

Estimating probabilities (Laplace estimation)

- The probabilities estimate by the fraction of times the event is observed to occur over the total number of opportunities. $\frac{n_c}{n}$
- It may provides poor/incorrect estimates when $\boldsymbol{n_c}$, is very small or zero.
- To avoid this difficulty we can adopt a Bayesian approach to estimating the probability, using the **m-estimate** defined as follows.

$$\frac{n_c + mp}{n + m}$$

assume uniform priors; if an attribute has ***k*** possible values

$$p = \frac{1}{k}$$

m is a constant called the ***equivalent sample size***

Text classification

- Classify e-mails
  - Y = {Spam, NotSpam}
- Classify news articles
  - Y = {what is the topic of the article?}
- Classify webpages
  - Y = {Student, professor, project, …}

- What about the features **X**?
  - The text!

Features X are entire document – X_i for i^{th} word in article

Article from rec.sport.hockey

```
Path: cantaloupe.srv.cs.cmu.edu!das-news.harvard.e
From: xxx@yyy.zzz.edu (John Doe)
Subject: Re: This year's biggest and worst (opinic
Date: 5 Apr 93 09:53:39 GMT

I can only comment on the Kings, but the most
obvious candidate for pleasant surprise is Alex
Zhitnik. He came highly touted as a defensive
defenseman, but he's clearly much more than that.
Great skater and hard shot (though wish he were
more accurate). In fact, he pretty much allowed
the Kings to trade away that huge defensive
liability Paul Coffey. Kelly Hrudey is only the
biggest disappointment if you thought he was any
good to begin with. But, at best, he's only a
mediocre goaltender. A better choice would be
Tomas Sandstrom, though not through any fault of
his own, but because some thugs in Toronto decided
```

NB for Text classification

- P(**X** | Y) is huge!!!
  - Article at least 1000 words, $\mathbf{X}=\{X_1,\ldots,X_{1000}\}$
  - X_i represents i^{th} word in document, i.e., the domain of X_i is entire vocabulary, e.g., Webster Dictionary (or more), 10,000 words, etc.

- NB assumption helps a lot!!!
  - $P(X_i=x_i \mid Y=y)$ is just the probability of observing word x_i in a document on topic y

$$h_{NB}(\mathbf{x}) = \arg\max_y P(y) \prod_{i=1}^{LengthDoc} P(x_i|y)$$

Bag of Words model

- Typical additional assumption:
  Position in document doesn't matter:
  $P(X_i=a \mid Y=y) = P(X_k=a \mid Y=y)$
  - "Bag of words" model – order of words on the page ignored
  - Sounds really silly, but often works very well!

$$P(y) \prod_{i=1}^{LengthDoc} P(x_i|y)$$

When the lecture is over, remember to wake up the person sitting next to you in the lecture room.

Bag of Words model

- Typical additional assumption:
  Position in document doesn't matter:
  $P(X_i=x_i \mid Y=y) = P(X_k=x_i \mid Y=y)$
  - "Bag of words" model – order of words on the page ignored
  - Sounds really silly, but often works very well!

$$P(y) \prod_{i=1}^{LengthDoc} P(x_i|y)$$

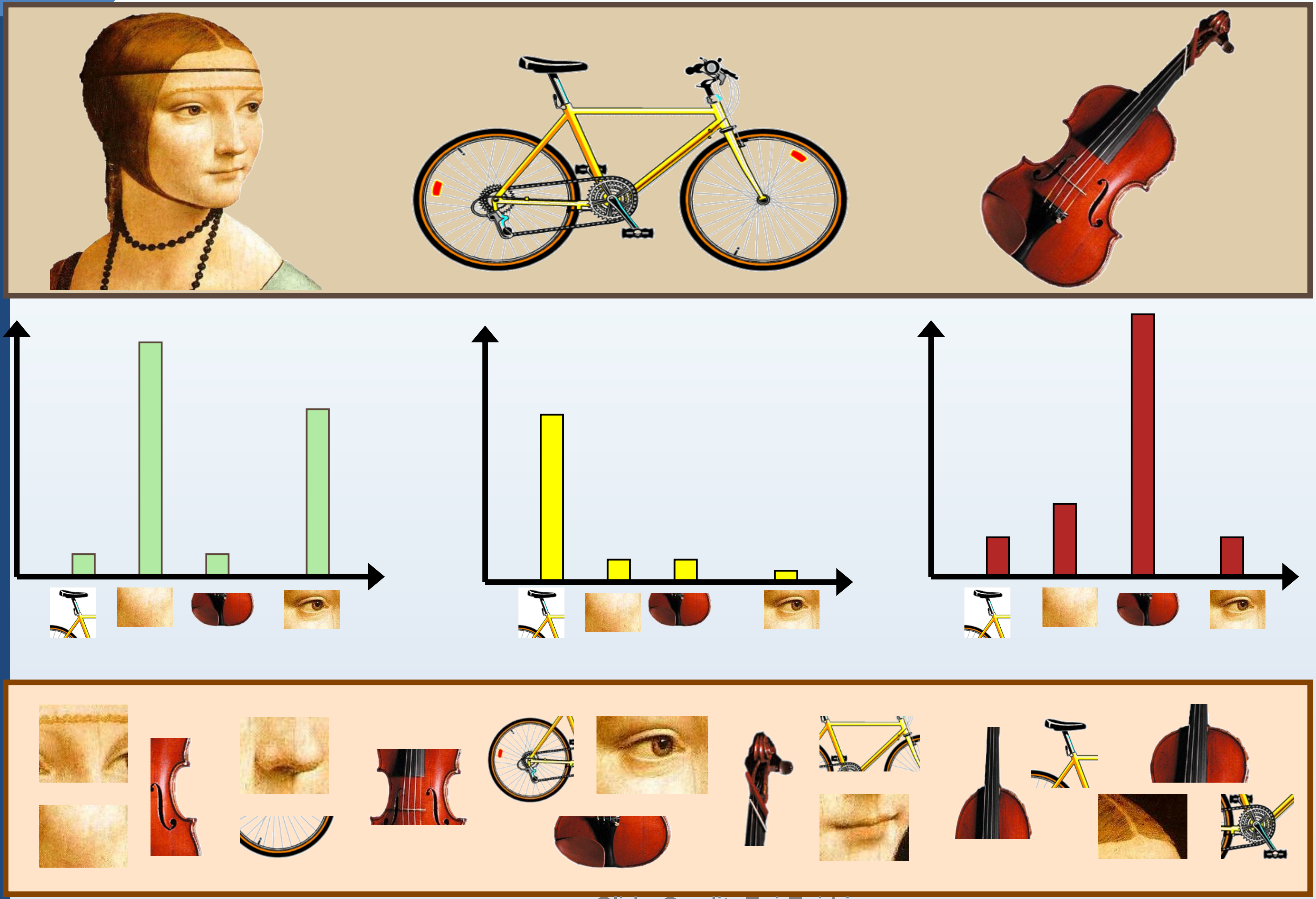
in is lecture lecture next over person remember room sitting the the the to to up wake when you

Bag of Words model

aardvark	0
about	2
all	2
Africa	1
apple	0
anxious	0
...	
gas	1
...	
oil	1
…	
Zaire	0

Object → Bag of 'words'

Slide Credit: Fei Fei Li



Slide Credit: Fei Fei Li

Example

$P(\text{Class} \mid \text{Doc}_5)=?$

$P(\text{Class} \mid \text{Doc}_5)=P(W_{\text{Doc5}} \mid \text{Class})\times P(\text{Class})$

	Doc	Words	Class
Training	1	Chinese Beijing Chinese	c
	2	Chinese Chinese Shanghai	c
	3	Chinese Macao	c
	4	Tokyo Japan Chinese	j
Test	5	Chinese Chinese Chinese Tokyo Japan	?

Prior:

$$P(Class = c) = \frac{3}{4}$$

$$P(Class = j) = \frac{1}{4}$$

Likelihood: $P(w_j|v_i) = \dfrac{count(w_j,v_i)+1}{count(v_i)+m}$

$P(Chinese|c)=(5+1)/(8+6)=3/7$
$P(Tokyo|c)=(0+1)/(8+6)=1/14$
$P(Japan|c)=(0+1)/(8+6)=1/14$

$P(Chinese|j)=(1+1)/(3+6)=2/9$
$P(Tokyo|j)=(1+1)/(3+6)=2/9$
$P(Japan|j)=(1+1)/(3+6)=2/9$

$$P(c|d_5) = P(Chinese|c)\times P(Chinese|c)\times P(Chinese|c)\times P(Tokyo|c)\times P(Japan|c)\times P(c)$$
$$=(3/7)\times(3/7)\times(3/7)\times(1/14)\times(1/14)\times(3/4) \approx 0.0003$$

$$P(j|d_5) = P(Chinese|j)\times P(Chinese|j)\times P(Chinese|j)\times P(Tokyo|j)\times P(Japan|j)\times P(j)$$
$$=(2/9)\times(2/9)\times(2/9)\times(2/9)\times(2/9)\times(1/4) \approx 0.0001$$

Technical Detail: Underflow

- Multiplying lots of probabilities, which are between 0 and 1 by definition, can result in floating-point underflow.
- Since $\log(xy) = \log(x) + \log(y)$, it is better to perform all computations by summing logs of probabilities rather than multiplying probabilities.
- Class with highest final un-normalized log probability
- score is still the most probable

$$c_{NB} = \underset{v_j \in V}{argmax}\left(logP(v_j) + \sum_{i=1}^{n} logP(a_i|v_j)\right)$$

What you need to know about NB

- Optimal decision using Bayes Classifier
  - the attribute values are conditionally independent given the target value
- Along with decision tree, neural networks, nearest neighbor, one of the most practical learning methods.
- When to use
  - Moderate or large training set available
  - Attributes that describe instances are conditionally independent given classification
- Very fast
  - Learns with one pass over the data
  - Testing linear in the number of attributes and of documents
  - Low storage requirements
- Successful applications
  - Diagnosis
  - Text classification (Bag of words model)
- Gaussian NB
  - Features are still conditionally independent
  - Each feature has a Gaussian distribution given class

Reading

- T. Mitchel, **Machine learning**, 1998. (ch. 6)
- C. M. Bishop, **Pattern recognition and machine learning**, Springer, 2006. (ch. 1)
- E. Alpaydin, **Introduction to Machine Learning**, 3rd ed., The MIT Press, 2014. (ch. 3)