ECE 5984: Introduction to Machine Learning

Topics:

- Classification: Naïve Bayes

Readings: Barber 10.1-10.3

Dhruv Batra
Virginia Tech

Administrativia

- HW2
  - Due: Friday ~~03/06~~, 03/15, 11:55pm
  - Implement linear regression, Naïve Bayes, Logistic Regression

- Need a couple of catch-up lectures
  - How about 4-6pm?

Administrativia

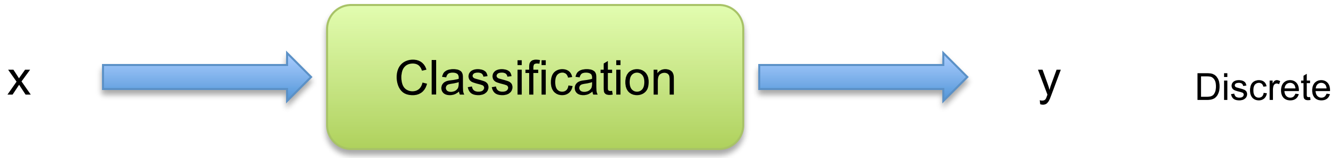
- Mid-term
  - When: March 18, class timing
  - Where: In class

  - Format: Pen-and-paper.
  - Open-book, open-notes, closed-internet.
    - No sharing.

  - What to expect: mix of
    - Multiple Choice or True/False questions
    - “Prove this statement”
    - “What would happen for this dataset?”

  - Material
    - Everything from beginning to class to (including) SVMs

New Topic:
Naïve Bayes
(your first probabilistic classifier)



Error Decomposition

- Approximation/Modeling Error
  - You approximated reality with model
- Estimation Error
  - You tried to learn model with finite data
- Optimization Error
  - You were lazy and couldn't/didn't optimize to completion
- Bayes Error
  - Reality just sucks
  - http://psych.hanover.edu/JavaTest/SDT/ROC.html

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?**

Slide Credit: Carlos Guestrin

Optimal classification

- **Theorem:** Bayes classifier h$_{Bayes}$ is optimal!
  - That is $error_{true}(h_{Bayes})) \leq error_{true}(h), \ \ \forall h(\mathbf{x})$

- **Proof**:

$$p(error_h) = \int_x p(error_h|x)p(x)dx$$

Slide Credit: Carlos Guestrin

Generative vs. Discriminative

- Using Bayes rule, optimal classifier

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

- 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)

Generative vs. Discriminative

- Generative Approach
  - Assume some functional form for P(X|Y), P(Y)
  - Estimate p(X|Y) and p(Y)
  - Use Bayes Rule to calculate P(Y| X=x)
  - Indirect computation of P(Y|X) through Bayes rule
  - But, **can generate a sample**, $P(X) = \sum_y P(y) P(X|y)$

- Discriminative Approach
  - Estimate p(y|x) directly OR
  - Learn “discriminant” function h(x)
  - Direct but cannot obtain a sample of the data, because P(X) is not available

Generative vs. Discriminative

- Generative:
  - Today: Naïve Bayes

- Discriminative:
  - Next: Logistic Regression

- NB & LR related to each other.

How hard is it to learn the optimal classifier?

- Categorical Data

- How do we represent these? How many parameters?
  - Class-Prior, P(Y):
    - Suppose Y is composed of *k* classes
  - Likelihood, P(**X**|Y):
    - Suppose **X** is composed of *d* binary features

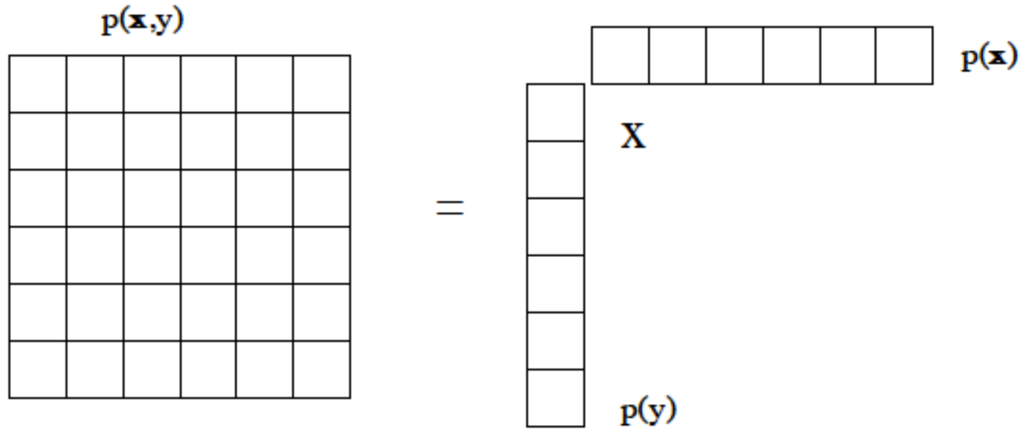
- Complex model → High variance with limited data!!!

Slide Credit: Carlos Guestrin

Independence to the rescue

- Two variables are independent iff their joint factors:

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



- Two variables are conditionally independent given a third one if for all values of the conditioning variable, the resulting slice factors:

$$p(x, y|z) = p(x|z)p(y|z) \qquad \forall z$$

The Naïve Bayes assumption

- Naïve Bayes assumption:
  - Features are independent given class:

$$\begin{aligned} P(X_1, X_2|Y) &= P(X_1|X_2, Y)P(X_2|Y) \\ &= P(X_1|Y)P(X_2|Y) \end{aligned}$$

  - More generally:

$$\boxed{P(X_1 ... X_d|Y) = \prod_i P(X_i|Y)}$$

- How many parameters now?
  - Suppose **X** is composed of *d* binary features

Slide Credit: Carlos Guestrin

The Naïve Bayes Classifier

- Given:
  - Class-Prior P(Y)
  - *d* conditionally independent features **X** given the class Y
  - For each X_i, we have likelihood $P(X_i|Y)$

- Decision rule:

$$\begin{aligned} y^* = h_{NB}(\mathbf{x}) &= \arg\max_y P(y) P(x_1, \ldots, x_n \mid y) \\ &= \arg\max_y P(y) \prod_i P(x_i|y) \end{aligned}$$

- If assumption holds, NB is optimal classifier!

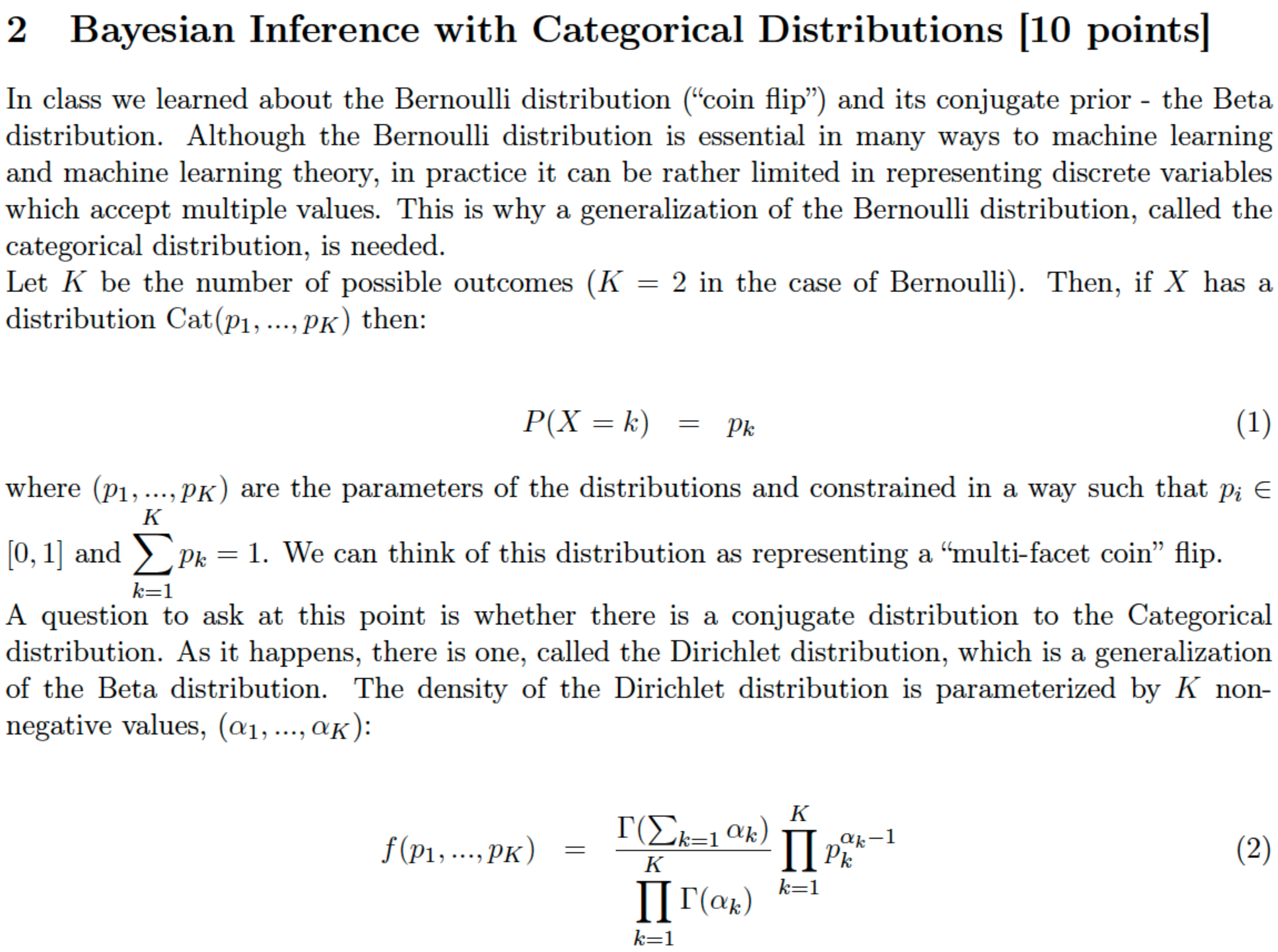
Slide Credit: Carlos Guestrin

MLE for the parameters of NB

- Given dataset
  - Count(A=a,B=b) #number of examples where A=a and B=b

- MLE for NB, simply:
  - Class-Prior: P(Y=y) =
  - Likelihood: P(X_i=x_i | Y=y) =

HW1



2 Bayesian Inference with Categorical Distributions [10 points]

In class we learned about the Bernoulli distribution ("coin flip") and its conjugate prior - the Beta distribution. Although the Bernoulli distribution is essential in many ways to machine learning and machine learning theory, in practice it can be rather limited in representing discrete variables which accept multiple values. This is why a generalization of the Bernoulli distribution, called the categorical distribution, is needed.

Let K be the number of possible outcomes ($K = 2$ in the case of Bernoulli). Then, if X has a distribution $\mathrm{Cat}(p_1, ..., p_K)$ then:

$$P(X = k) \quad = \quad p_k \tag{1}$$

where $(p_1, ..., p_K)$ are the parameters of the distributions and constrained in a way such that $p_i \in [0, 1]$ and $\sum_{k=1}^{K} p_k = 1$. We can think of this distribution as representing a "multi-facet coin" flip.

A question to ask at this point is whether there is a conjugate distribution to the Categorical distribution. As it happens, there is one, called the Dirichlet distribution, which is a generalization of the Beta distribution. The density of the Dirichlet distribution is parameterized by K non-negative values, $(\alpha_1, ..., \alpha_K)$:

$$f(p_1, ..., p_K) \quad = \quad \frac{\Gamma(\sum_{k=1} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} p_k^{\alpha_k - 1} \tag{2}$$

Naïve Bayes

- In class demo

- Variables
  - Y = {did-not-watch superbowl, watched superbowl}
  - X_1 = {domestic, international}
  - X_2 = {<2 years at VT, >=2years}

- Estimate
  - P(Y=1)
  - $P(X_1=0 \mid Y=0)$, $P(X_2=0 \mid Y=0)$
  - $P(X_1=0 \mid Y=1)$, $P(X_2=0 \mid Y=1)$

- Prediction: argmax_y $P(Y=y)P(x_1|Y=y)P(x_2|Y=y)$

Subtleties of NB classifier 1 – Violating the NB assumption

- Usually, features are not conditionally independent:

$$\boxed{P(X_1 ... X_d|Y) \neq \prod_i P(X_i|Y)}$$

- 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

Slide Credit: Carlos Guestrin

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???

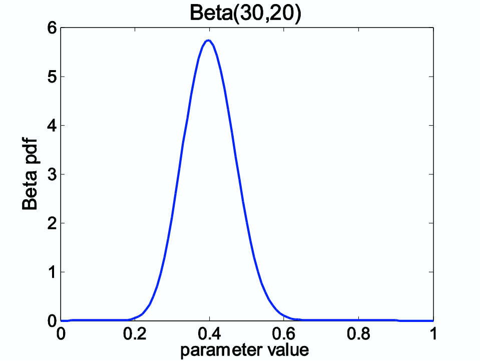
Slide Credit: Carlos Guestrin

Recall MAP for Bernoulli-Beta

$$P(\theta \mid \mathcal{D}) = \frac{\theta^{\beta_H+\alpha_H-1}(1-\theta)^{\beta_T+\alpha_T-1}}{B(\beta_H+\alpha_H,\beta_T+\alpha_T)} \sim Beta(\beta_H+\alpha_H,\beta_T+\alpha_T)$$

- MAP: use most likely parameter:

$$\widehat{\theta} = \arg\max_{\theta} P(\theta \mid \mathcal{D}) =$$



- Beta prior equivalent to extra flips

Slide Credit: Carlos Guestrin

Bayesian learning for NB parameters – a.k.a. smoothing

- Prior on parameters
  - Dirichlet everything!

- MAP estimate

- Now, even if you never observe a feature/class, posterior probability never zero

Slide Credit: Carlos Guestrin

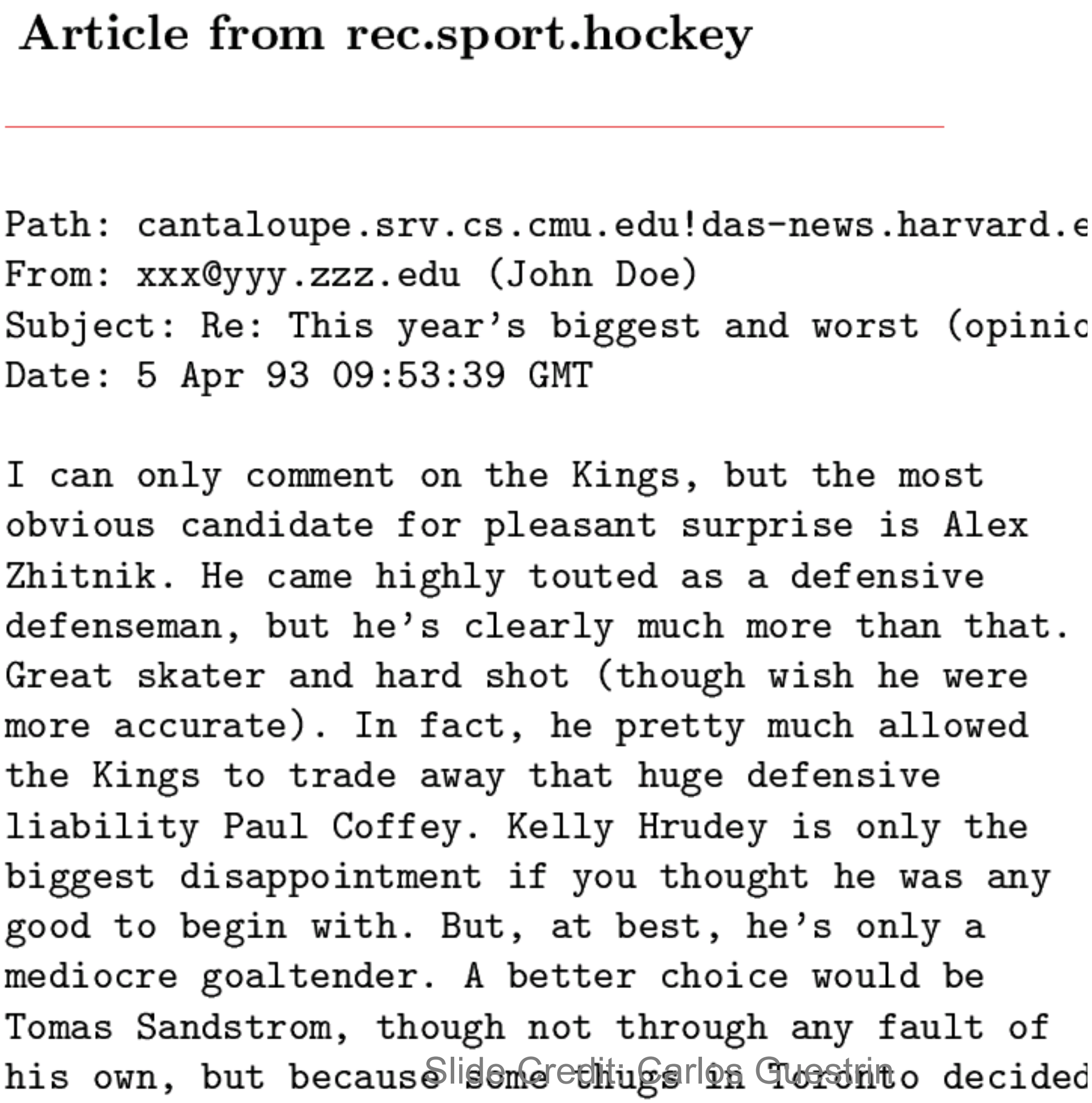
Text classification

- 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!

Slide Credit: Carlos Guestrin

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

Slide Credit: Carlos Guestrin

NB for Text classification

- P(**X**|Y) is huge!!!
  - Article at least 1000 words, **X**={$X_1,\ldots,X_{1000}$}
  - X_i represents ith 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|Y=y) is just the probability of observing word x_i in a document on topic y

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

(C) Dhruv Batra

Slide Credit: Carlos Guestrin

Bag of Words model

- Typical additional assumption:
  Position in document doesn't matter:
  $P(X_i=a|Y=y) = P(X_k=a|Y=y)$
  - "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.

Slide Credit: Carlos Guestrin

Bag of Words model

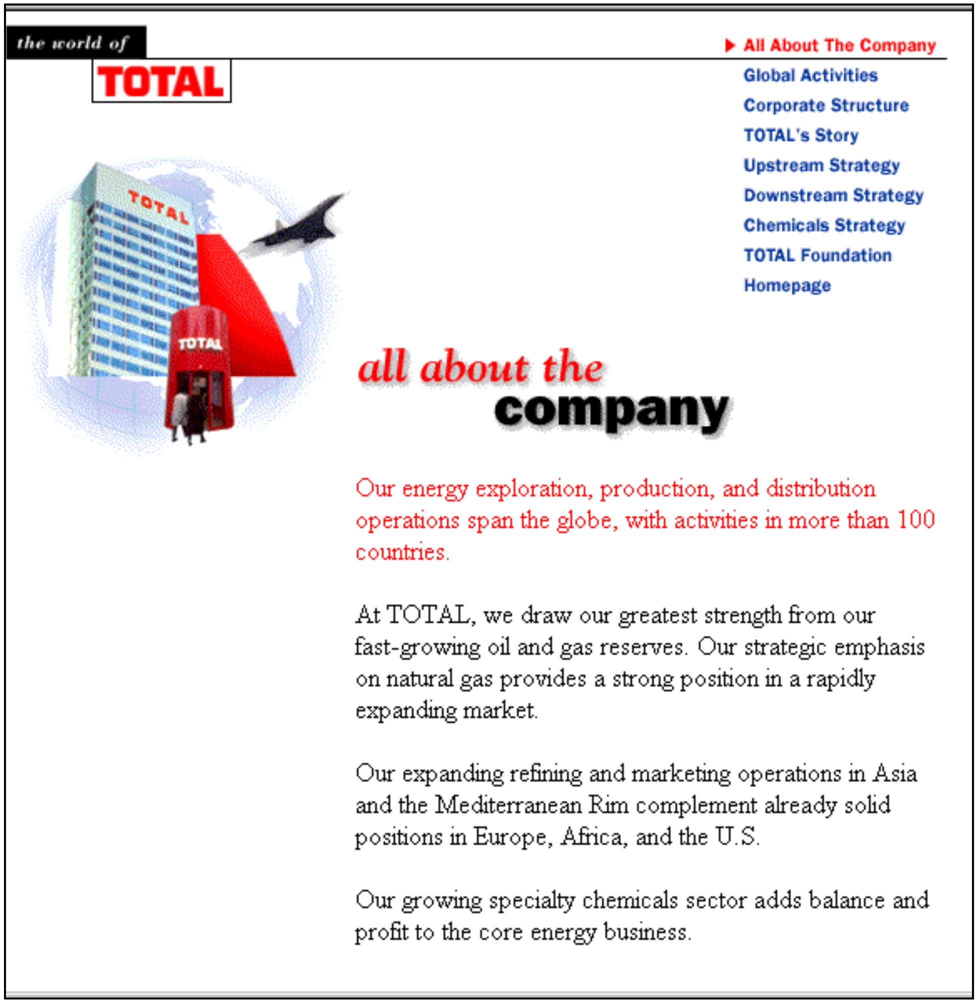
- Typical additional assumption:
  Position in document doesn't matter:
  $P(X_i=x_i|Y=y) = P(X_k=x_i|Y=y)$
  - "Bag of words" model – order of words on the page ignored
  - Sounds really silly, but often works very well!

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

in is lecture lecture next over person remember room

sitting the the the to to up wake when you

Bag of Words model



all about the
company

Our energy exploration, production, and distribution operations span the globe, with activities in more than 100 countries.

At TOTAL, we draw our greatest strength from our fast-growing oil and gas reserves. Our strategic emphasis on natural gas provides a strong position in a rapidly expanding market.

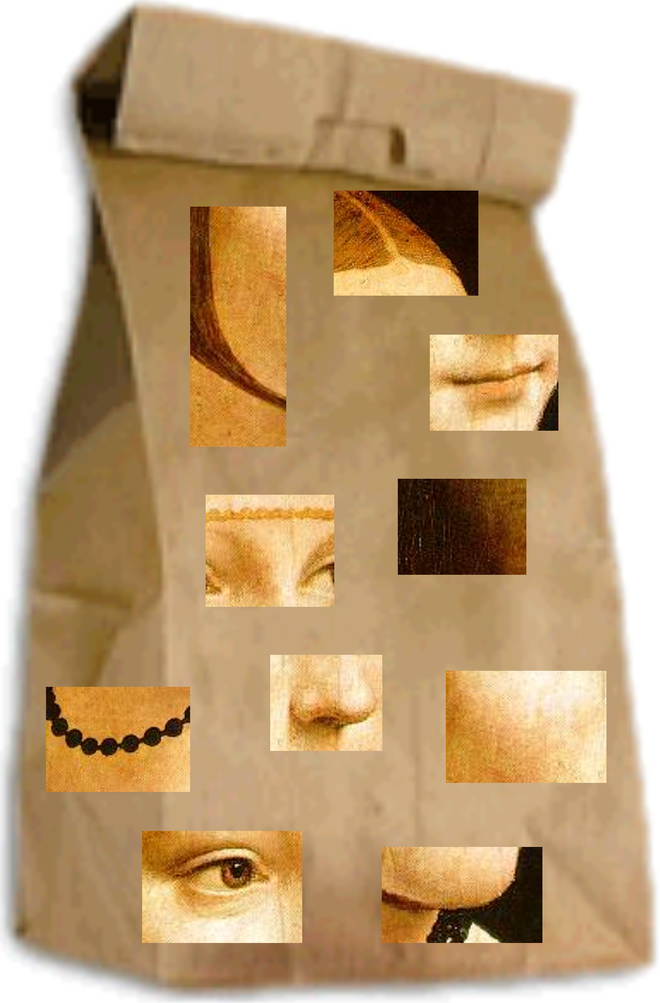
Our expanding refining and marketing operations in Asia and the Mediterranean Rim complement already solid positions in Europe, Africa, and the U.S.

Our growing specialty chemicals sector adds balance and profit to the core energy business.

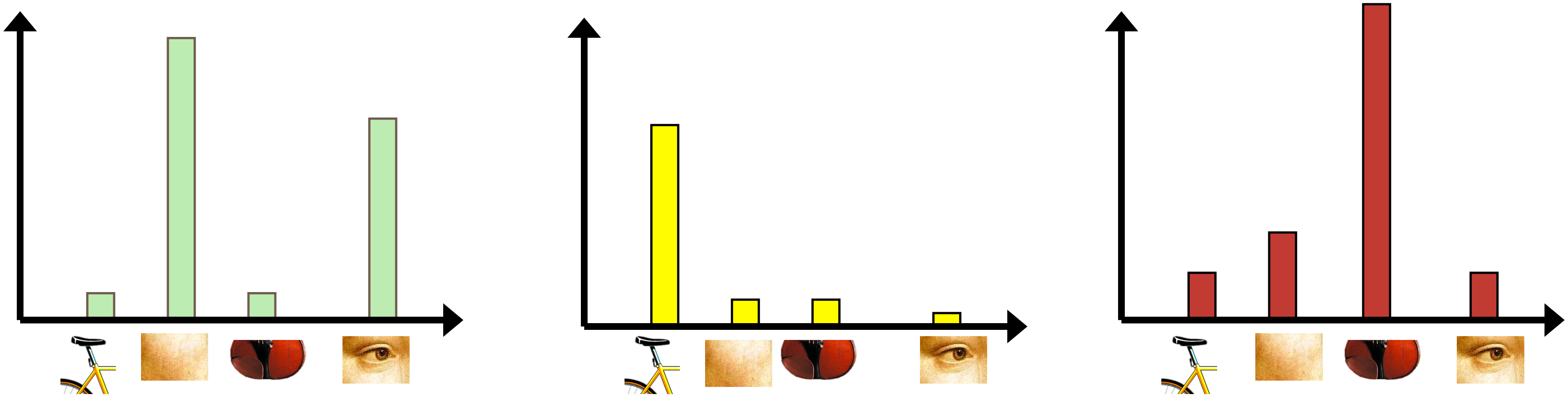
aardvark	0
about	2
all	2
Africa	1
apple	0
anxious	0
...	
gas	1
...	
oil	1
…	
Zaire	0

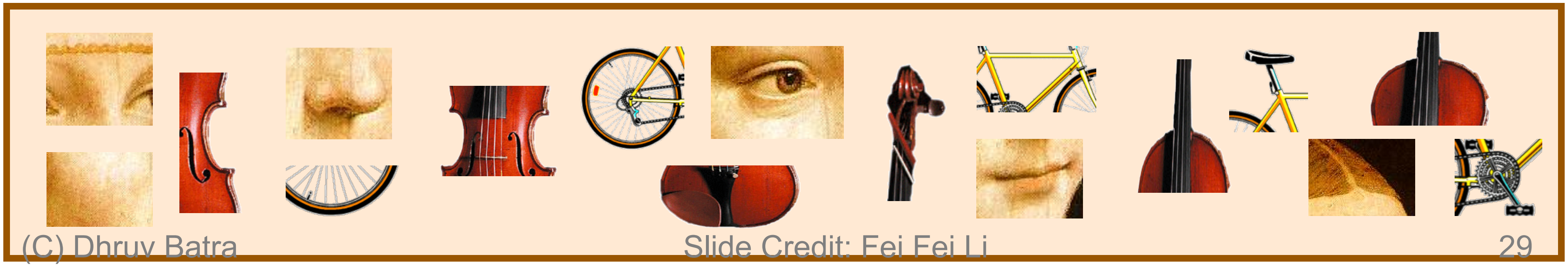
Slide Credit: Carlos Guestrin

Object → **Bag of ‘words’**



Slide Credit: Fei Fei Li

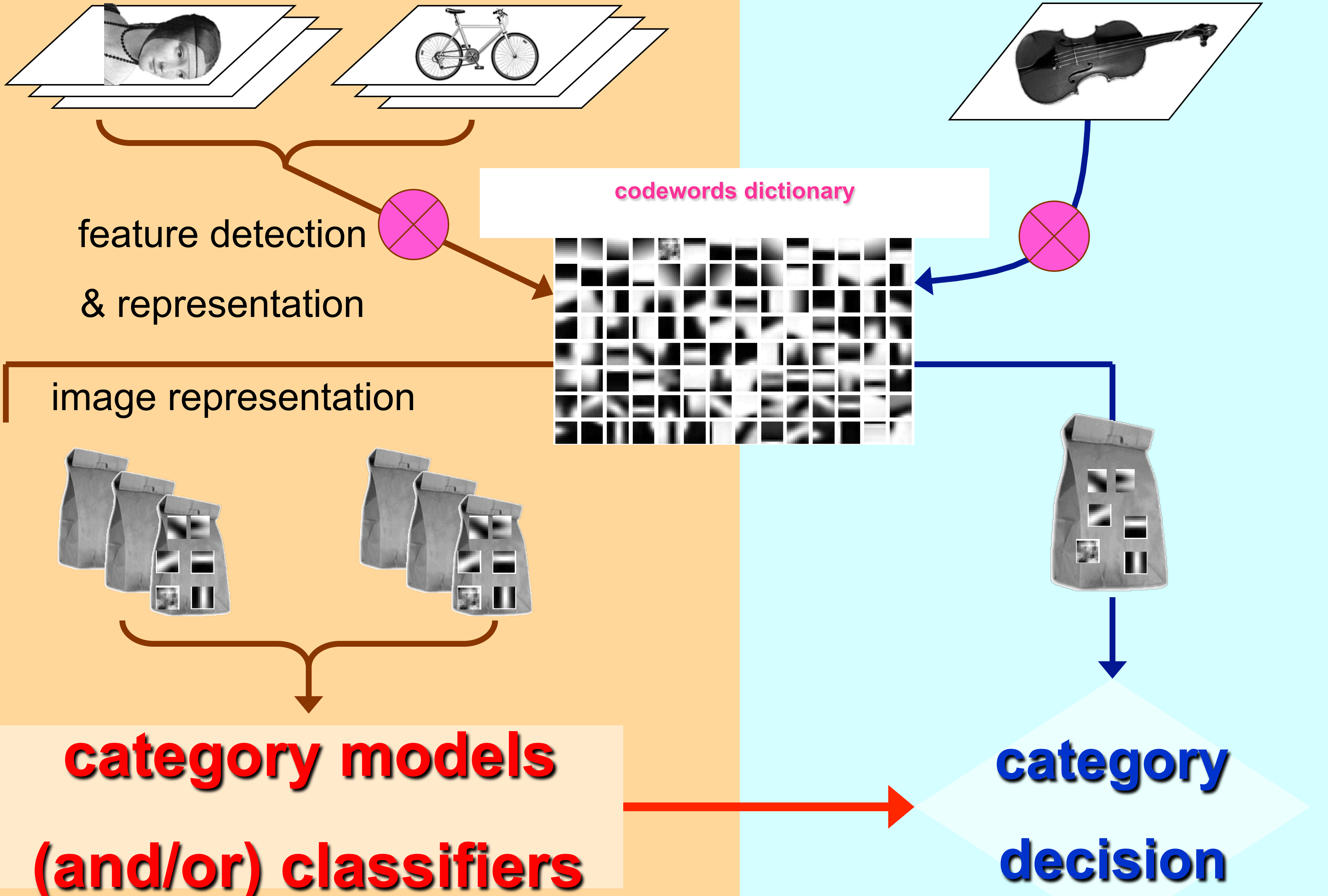




Slide Credit: Fei Fei Li

learning

recognition



codewords dictionary

feature detection

& representation

image representation

category models

(and/or) classifiers

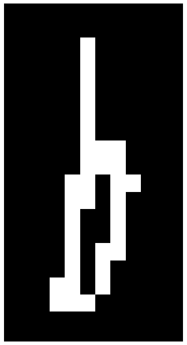
category

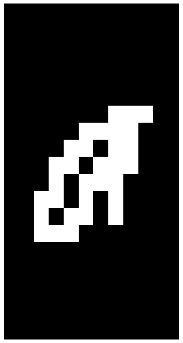
decision

Slide Credit: Fei Fei Li

What if we have continuous X_i ?

Eg., character recognition: X_i is i^th pixel





Gaussian Naïve Bayes (GNB):

$$P(X_i = x \mid Y = y_k) = \frac{1}{\sigma_{ik}\sqrt{2\pi}} \; e^{\frac{-(x-\mu_{ik})^2}{2\sigma_{ik}^2}}$$

Sometimes assume variance

- is independent of Y (i.e., σ_i),
- or independent of X_i (i.e., σ_k)
- or both (i.e., σ)

Slide Credit: Carlos Guestrin

Estimating Parameters: *Y* discrete, *X_i* continuous

Maximum likelihood estimates:

$$\widehat{\mu}_{ik} =$$

$$\widehat{\sigma}_{ik}^2 =$$

What you need to know about NB

- Optimal decision using Bayes Classifier

- Naïve Bayes classifier
  - What's the assumption
  - Why we use it
  - How do we learn it
  - Why is Bayesian estimation of NB parameters important

- Text classification
  - Bag of words model

- Gaussian NB
  - Features are still conditionally independent
  - Each feature has a Gaussian distribution given class

Generative vs. Discriminative

- Using Bayes rule, optimal classifier

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

- 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)

Today: Logistic Regression

- Main idea
  - Think about a 2 class problem {0,1}
  - Can we regress to P(Y=1 | X=x)?

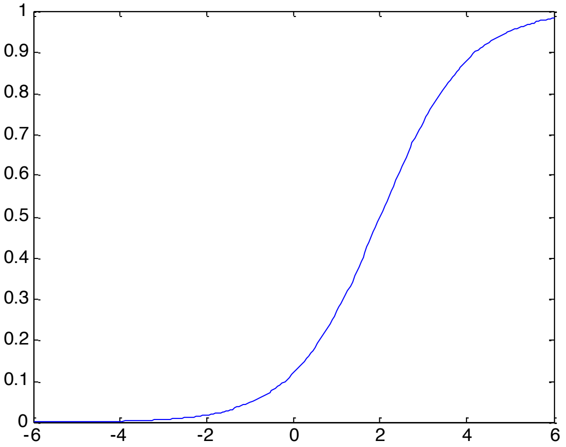
- Meet the Logistic or Sigmoid function
  - Crunches real numbers down to 0-1

- Model
  - In regression: $y \sim N(w'x, \lambda^2)$
  - Logistic Regression: $y \sim \text{Bernoulli}(\sigma(w'x))$

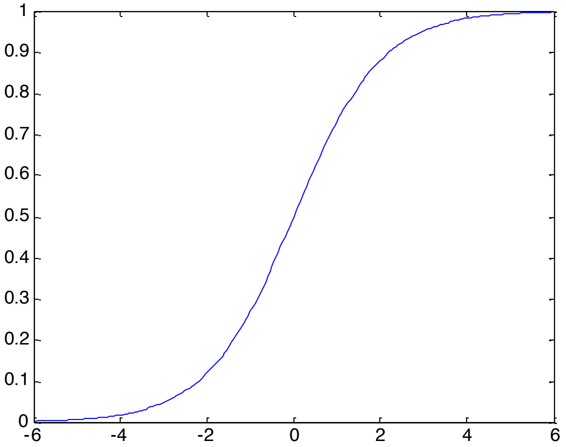
Understanding the sigmoid

$$\sigma(w_0 + \sum_i w_i x_i) = \frac{1}{1 + e^{-w_0 - \sum_i w_i x_i}}$$

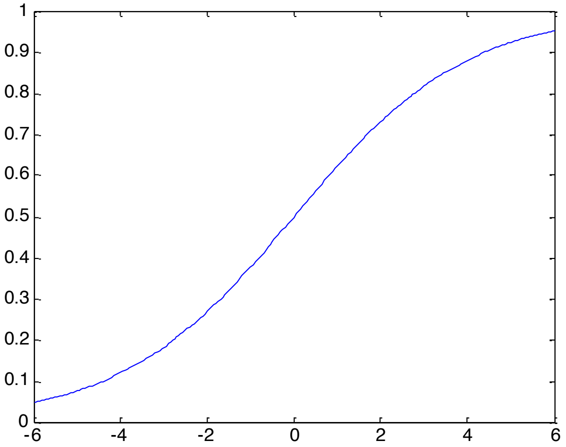
w_0=2, w_1=1



w_0=0, w_1=1



w_0=0, w_1=0.5



Slide Credit: Carlos Guestrin