# Bayesian Classifiers, Conditional Independence and Naïve Bayes

Required reading:

- Mitchell draft chapter, sections 1 and 2. (available on class website)

Machine Learning 10-601

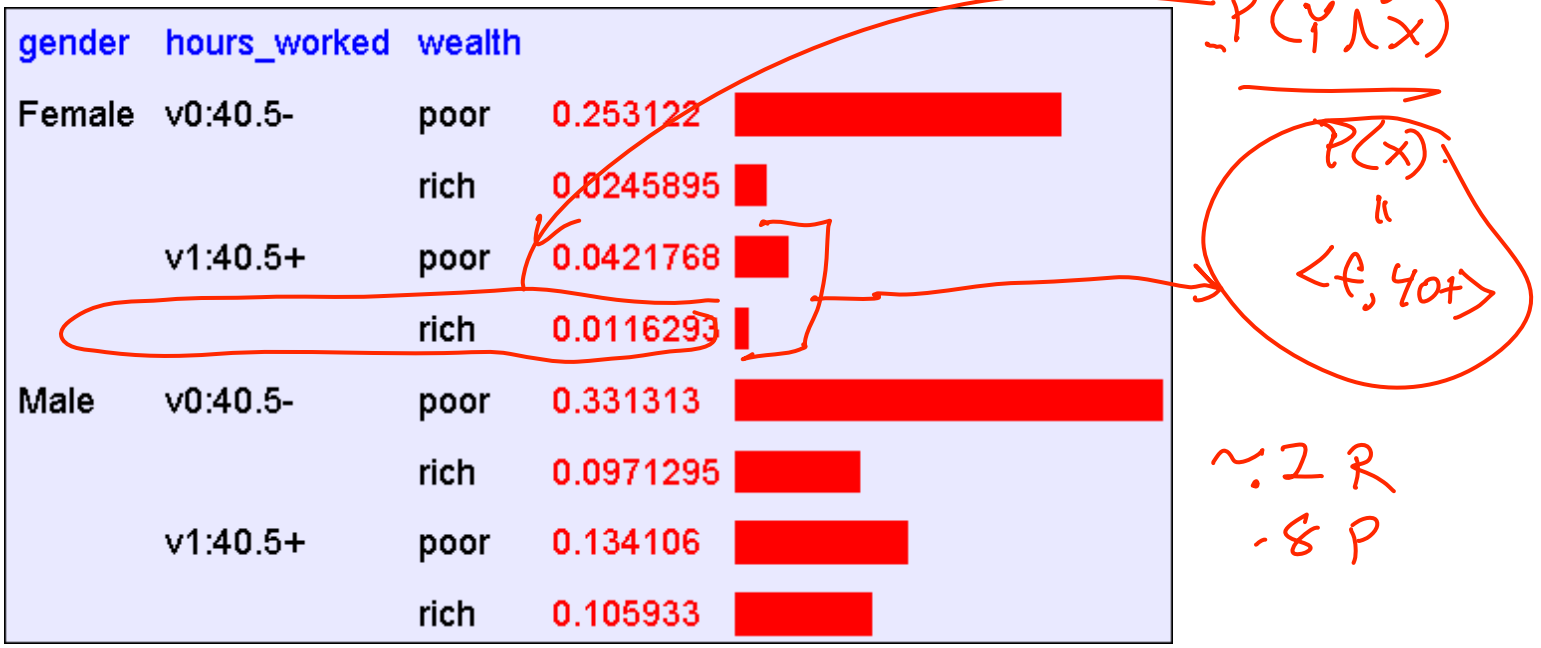
Tom M. Mitchell
Machine Learning Department
Carnegie Mellon University

Jan 28, 2009
Feb 2, 2009

# Let's learn classifiers by learning P(Y|X)

Suppose Y=wealth, X=<gender, hours_worked>

| gender | hours_worked | wealth | |
|---|---|---|---|
| Female | v0:40.5- | poor | 0.253122 |
| | | rich | 0.0245895 |
| | v1:40.5+ | poor | 0.0421768 |
| | | rich | 0.0116293 |
| Male | v0:40.5- | poor | 0.331313 |
| | | rich | 0.0971295 |
| | v1:40.5+ | poor | 0.134106 |
| | | rich | 0.105933 |



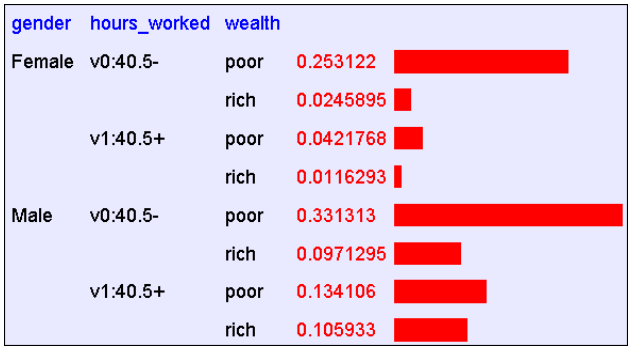

n Boolean vars → $2^n$ rows

$2^N$ P(Y,X) terms

# How many parameters must we estimate?

Suppose X =<X1,… Xn>

where Xi and Y are boolean RV's



| gender | hours_worked | wealth | |
|---|---|---|---|
| Female | v0:40.5- | poor | 0.253122 |
| | | rich | 0.0245895 |
| | v1:40.5+ | poor | 0.0421768 |
| | | rich | 0.0116293 |
| Male | v0:40.5- | poor | 0.331313 |
| | | rich | 0.0971295 |
| | v1:40.5+ | poor | 0.134106 |
| | | rich | 0.105933 |

To estimate P(Y| X) = P(Y| X1, X2, … Xn)

# Can we reduce params by using Bayes Rule?

Suppose X =<X1,… Xn>

where Xi and Y are boolean RV's

$2^n$ params

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

$P(X_1, X_2 \cdots X_n | Y)$

$P(X_1=1, X_2=0, \cdots X_n=1 | Y=1)$

$2^{n-1}$ $\quad$ 2

$2 \times (2^{n-1}) - 1$

$2^n - 2$ params for numerator

$+ P(Y=1)$

$2^n - 1$

$\sum_j P(X|Y=j)P(Y=j)$

$P(X=1|Y=1) = .7$

$P(X=0|Y=1) = .3$

$P(X=1|Y=0) = .1$

$P(X=0|Y=0) = .89$

$2^n$

$2n$

# Bayes Rule

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

Which is shorthand for:

$$(\forall i,j) P(Y = y_i | X = x_j) = \frac{P(X = x_j | Y = y_i) P(Y = y_i)}{P(X = x_j)}$$

Equivalently:

$$(\forall i,j) P(Y = y_i | X = x_j) = \frac{P(X = x_j | Y = y_i) P(Y = y_i)}{\sum_k P(X = x_j | Y = y_k) P(Y = y_k)}$$

# Naïve Bayes

Naïve Bayes assumes

$$P(X_1 \ldots X_n|Y) = \prod_i P(X_i|Y)$$

i.e., that $X_i$ and $X_j$ are conditionally independent given Y, for all $i \neq j$

# Conditional Independence

Definition: X is conditionally independent of Y given Z, if the probability distribution governing X is independent of the value of Y, given the value of Z

$$(\forall i, j, k) P(X = x_i | Y = y_j, Z = z_k) = P(X = x_i | Z = z_k)$$

Which we often write

$$P(X|Y, Z) = P(X|Z)$$

E.g.,

$$P(Thunder|Rain, Lightning) = P(Thunder|Lightning)$$

Naïve Bayes uses assumption that the $X_i$ are conditionally independent, given Y

$P(Y=1) = \theta$

Given this assumption, then:

$$P(X_1, X_2|Y) = P(X_1|X_2, Y)P(X_2|Y)$$
$$= P(X_1|Y)P(X_2|Y)$$

Cond Ind. assump.

in general: $P(X_1 \ldots X_n|Y) = \prod_i P(X_i|Y)$

$P(X_3=0|Y=0) + P(X_3=1|Y=0) = 1$ $\quad (n-1) \times 2$ $\quad n, 1$

How many parameters needed to describe P(X|Y)? P(Y)?

- Without conditional indep assumption?
- With conditional indep assumption? $2n-2$

# How many parameters to estimate?

P(X1, ... Xn | Y), all variables boolean

Without conditional independence assumption:

With conditional independence assumption:

# Naïve Bayes in a Nutshell

Bayes rule:

$$P(Y = y_k|X_1 \dots X_n) = \frac{P(Y = y_k)P(X_1 \dots X_n|Y = y_k)}{\sum_j P(Y = y_j)P(X_1 \dots X_n|Y = y_j)}$$

Assuming conditional independence among $X_i$'s:

$$P(Y = y_k|X_1 \dots X_n) = \frac{P(Y = y_k)\prod_i P(X_i|Y = y_k)}{\sum_j P(Y = y_j)\prod_i P(X_i|Y = y_j)}$$

So, classification rule for $X^{new} = < X_1, \dots, X_n >$ is:

$$Y^{new} \leftarrow \arg\max_{y_k} \; P(Y = y_k)\prod_i P(X_i^{new}|Y = y_k)$$

$$P(Y=1|X_1 \cdots X_n) \geq P(Y=0|X_1 \cdots X_n)$$

# Naïve Bayes Algorithm – discrete $X_i$

- Train Naïve Bayes (examples)

  for each* value $y_k$

    estimate $\pi_k \equiv P(Y = y_k)$

    for each* value $x_{ij}$ of each attribute $X_i$

      estimate $\theta_{ijk} \equiv P(X_i = x_{ij} | Y = y_k)$

- Classify ($X^{new}$)

$$Y^{new} \leftarrow \arg\max_{y_k} P(Y = y_k) \prod_i P(X_i^{new} | Y = y_k)$$

$$Y^{new} \leftarrow \arg\max_{y_k} \pi_k \prod_i \theta_{ijk}$$

* probabilities must sum to 1, so need estimate only n-1 parameters...

# Estimating Parameters: $Y, X_i$ discrete-valued

Maximum likelihood estimates (MLE's):

$$\hat{\pi}_k = \hat{P}(Y = y_k) = \frac{\#D\{Y = y_k\}}{|D|}$$

$$\hat{\theta}_{ijk} = \hat{P}(X_i = x_{ij} | Y = y_k) = \frac{\#D\{X_i = x_{ij} \wedge Y = y_k\}}{\#D\{Y = y_k\}}$$

Number of items in D for which $Y=y_k$

# Example: Live in Sq Hill?

- S=1 iff live in Squirrel Hill
- G=1 iff shop at SH Giant Eagle
- D=1 iff Drive to CMU
- M=1 iff Rachel Maddow fan

$P(S|G,D,M) = \frac{P(G,D,M|S)\,P(S)}{P(G,D,M)} = \frac{P(G|S)\,P(D|S)\,P(M|S)\,P(S)}{P(G,D,M)}$

$P(S=1) = \frac{9}{9+33} = .21$ $\quad P(S=0) = P(S=1) - 1 = .79$

$P(G=1|S=1) = \frac{7}{9}$ $\quad P(G=0|S=1) = 1 - \frac{7}{9} = \frac{2}{9}$

$P(G=1|S=0) = \frac{15}{33}$ $\quad P(G=0|S=0) = 1 - \frac{15}{33} = \frac{18}{33}$

$P(D=1|S=1) = \frac{4}{9}$ $\quad P(D=0|S=1) = \frac{5}{9}$ $\quad G=0$

$P(D=1|S=0) = \frac{1}{33}$ $\quad P(D=0|S=0) = \frac{32}{33}$ $\quad D=0$

$P(M=1|S=1) = \frac{1}{9}$ $\quad P(M=0|S=1) = \frac{8}{9}$ $\quad M=0$

$P(M=1|S=0) = \frac{1}{33}$ $\quad P(M=0|S=0) = \frac{32}{33}$

$P(G=0|S=1)\,P(D=0|S=1)\,P(M=0|S=1)\,P(S=1) = 0.02$

$P(G=0|S=0)\,P(D=0|S=0)\,P(M=0|S=0)\,P(S=0) = 0.4$

$\frac{18}{33}$

# Example: Live in Sq Hill? P(S|G,D,M)

- S=1 iff live in Squirrel Hill
- G=1 iff shop at SH Giant Eagle
- D=1 iff Drive to CMU
- M=1 iff Rachel Maddow fan

# Example: Live in Sq Hill? P(S|G,D,M)

- S=1 iff live in Squirrel Hill
- G=1 iff shop at SH Giant Eagle
- D=1 iff Drive to CMU
- M=1 iff Rachel Maddow fan

# Naïve Bayes: Subtlety #1

$\theta_{MLE} = \arg\max_\theta P(D|\theta)$

$\theta_{MAP} = \arg\max_\theta P(\theta|D) \propto P(D|\theta)P(\theta)$

If unlucky, our MLE estimate for $P(X_i \mid Y)$ might be zero. (e.g., $X_{373}$= Birthday_Is_January_30_1990)

- Why worry about just one parameter out of many?

- What can be done to avoid this?

$$P(x) = \frac{\#D(x=1) + \beta_1}{\#D(x=1) + \#D(x=0) + \beta_0 + \beta_1}$$

MAP Est with ~~$P(\theta) \neq 0$~~

$P(\theta)$ is a Beta Prior

# Estimating Parameters: $Y, X_i$ discrete-valued

Maximum likelihood estimates:

$$\hat{\pi}_k = \hat{P}(Y = y_k) = \frac{\#D\{Y = y_k\}}{|D|}$$

$$\hat{\theta}_{ijk} = \hat{P}(X_i = x_{ij}|Y = y_k) = \frac{\#D\{X_i = x_{ij} \wedge Y = y_k\}}{\#D\{Y = y_k\}}$$

MAP estimates (Dirichlet priors):

$$\hat{\pi}_k = \hat{P}(Y = y_k) = \frac{\#D\{Y = y_k\} + \alpha_k}{|D| + \sum_m \alpha_m}$$

Only difference: "imaginary" examples

$$\hat{\theta}_{ijk} = \hat{P}(X_i = x_{ij}|Y = y_k) = \frac{\#D\{X_i = x_{ij} \wedge Y = y_k\} + \alpha'_k}{\#D\{Y = y_k\} + \sum_m \alpha'_m}$$

# Naïve Bayes: Subtlety #2

Ⓐ $\hat{P}(X_2|Y) = \hat{P}(X_2'|Y)$

Ⓑ $\hat{P}(X_2|Y) = 1 - P(X_2''|Y)$

Often the $X_i$ are not really conditionally independent

- We use Naïve Bayes in many cases anyway, and it often works pretty well
  - often the right classification, even when not the right probability (see [Domingos&Pazzani, 1996])

- What is effect on estimated P(Y|X)?
  - Special case: what if we add two copies: $X_i = X_k$

Ⓐ $X_2' \equiv X_2$

Ⓑ $X_2' \equiv 1 - X_2$

$P(Y|X_1, X_2) = P(Y)\,P(X_1|Y)\,P(X_2|Y)\,\frac{1}{2} \rightarrow X_2'$

$P(Y)\,P(X_1|Y)\,P(X_2|Y)\,P(X_2'|Y)\,\frac{1}{2^2}$

# Learning to classify text documents

- Classify which emails are spam?
- Classify which emails promise an attachment?
- Classify which web pages are student home pages?

How shall we represent text documents for Naïve Bayes?

# 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

# Learning to Classify Text

Target concept $Interesting? : Document \rightarrow \{+, -\}$

1. Represent each document by vector of words
   - one attribute per word position in document
2. Learning: Use training examples to estimate
   - $P(+)$
   - $P(-)$
   - $P(doc|+)$
   - $P(doc|-)$

Naive Bayes conditional independence assumption

$$P(doc|v_j) = \prod_{i=1}^{length(doc)} P(a_i = w_k|v_j)$$

where $P(a_i = w_k|v_j)$ is probability that word in position $i$ is $w_k$, given $v_j$

one more assumption:
$P(a_i = w_k|v_j) = P(a_m = w_k|v_j), \forall i, m$

# Baseline: Bag of Words Approach


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

# Twenty NewsGroups

Given 1000 training documents from each group
Learn to classify new documents according to
which newsgroup it came from

| | |
|---|---|
| comp.graphics | misc.forsale |
| comp.os.ms-windows.misc | rec.autos |
| comp.sys.ibm.pc.hardware | rec.motorcycles |
| comp.sys.mac.hardware | rec.sport.baseball |
| comp.windows.x | rec.sport.hockey |
| | |
| alt.atheism | sci.space |
| soc.religion.christian | sci.crypt |
| talk.religion.misc | sci.electronics |
| talk.politics.mideast | sci.med |
| talk.politics.misc | |
| talk.politics.guns | |

Naive Bayes: 89% classification accuracy

LEARN_NAIVE_BAYES_TEXT($Examples, V$)

1. *collect all words and other tokens that occur in Examples*

- $Vocabulary \leftarrow$ all distinct words and other tokens in $Examples$

2. *calculate the required $P(v_j)$ and $P(w_k|v_j)$ probability terms*

- For each target value $v_j$ in $V$ do
  - $docs_j \leftarrow$ subset of $Examples$ for which the target value is $v_j$
  - $P(v_j) \leftarrow \frac{|docs_j|}{|Examples|}$
  - $Text_j \leftarrow$ a single document created by concatenating all members of $docs_j$
  - $n \leftarrow$ total number of words in $Text_j$ (counting duplicate words multiple times)
  - for each word $w_k$ in $Vocabulary$
    - $n_k \leftarrow$ number of times word $w_k$ occurs in $Text_j$
    - $P(w_k|v_j) \leftarrow \frac{n_k+1}{n+|Vocabulary|}$

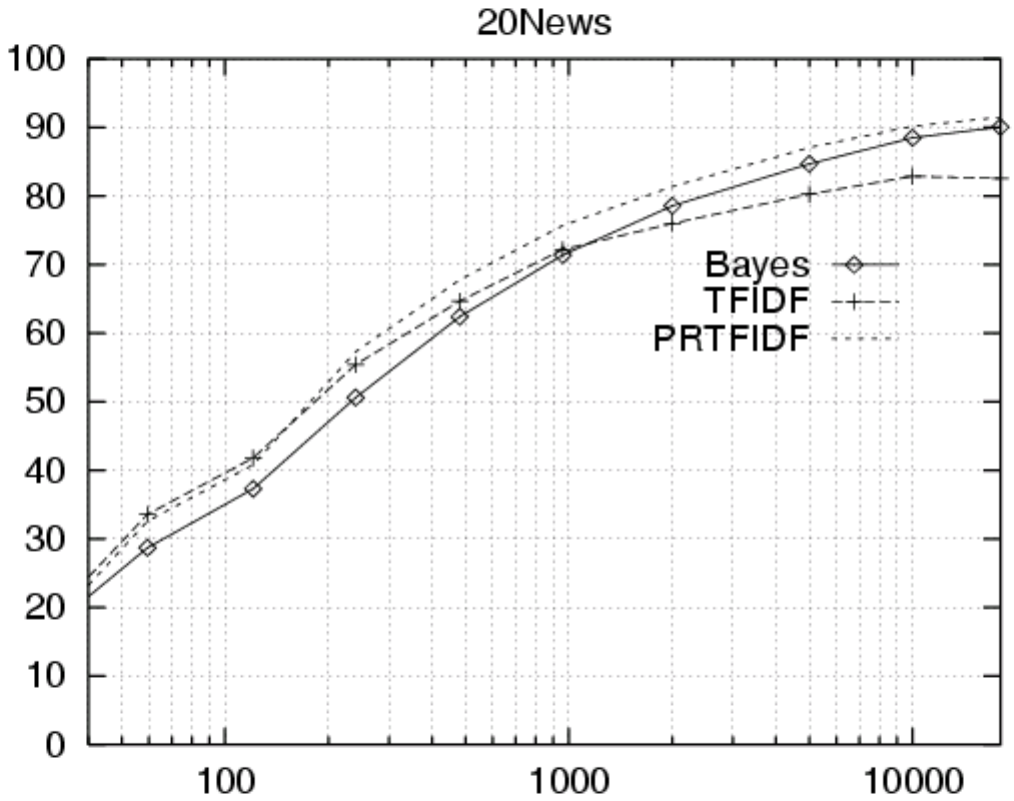
For code and data, see
www.cs.cmu.edu/~tom/mlbook.html
click on "Software and Data"

CLASSIFY_NAIVE_BAYES_TEXT($Doc$)

- $positions \leftarrow$ all word positions in $Doc$ that contain tokens found in $Vocabulary$
- Return $v_{NB}$, where

$$v_{NB} = \underset{v_j \in V}{\operatorname{argmax}} P(v_j) \prod_{i \in positions} P(a_i | v_j)$$

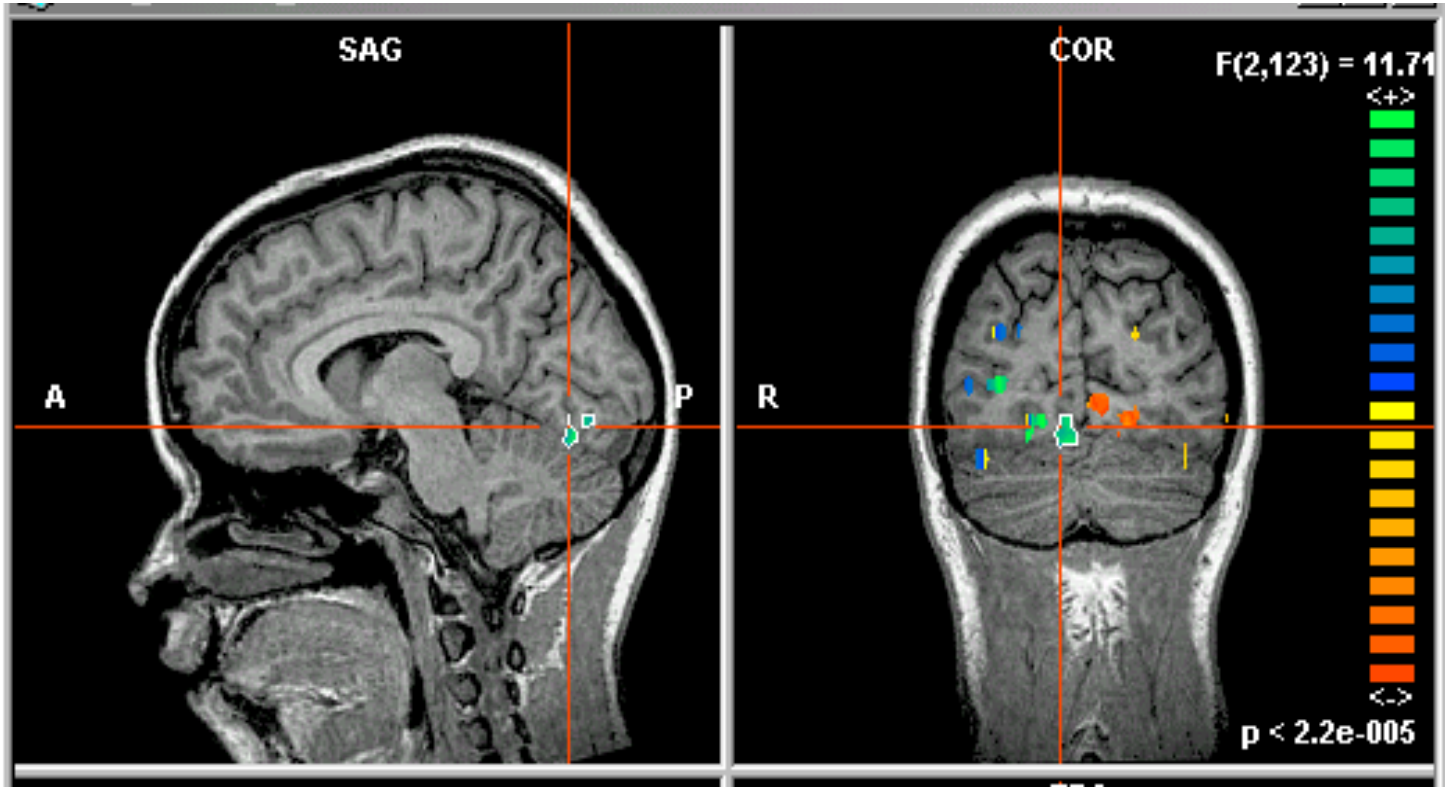
# Learning Curve for 20 Newsgroups




Accuracy vs. Training set size (1/3 withheld for test)

# What if we have continuous $X_i$ ?

Eg., image classification: $X_i$ is i$^{th}$ pixel



# What if we have continuous $X_i$ ?

Eg., image classification: $X_i$ is i$^{th}$ pixel

Gaussian Naïve Bayes (GNB): assume

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

Sometimes assume variance

- is independent of Y (i.e., $\sigma_i$),
- or independent of $X_i$ (i.e., $\sigma_k$)
- or both (i.e., $\sigma$)

# Gaussian (aka Normal) Distribution

# Gaussian Naïve Bayes Algorithm – continuous $X_i$ (but still discrete Y)

- Train Naïve Bayes (examples)

  for each value $y_k$

  estimate* $\pi_k \equiv P(Y = y_k)$

  for each attribute $X_i$ estimate

  class conditional mean $\mu_{ik}$, variance $\sigma_{ik}$

- Classify ($X^{new}$)

$$Y^{new} \leftarrow \arg\max_{y_k} P(Y = y_k) \prod_i P(X_i^{new} | Y = y_k)$$

$$Y^{new} \leftarrow \arg\max_{y_k} \pi_k \prod_i Normal(X_i^{new}, \mu_{ik}, \sigma_{ik})$$

* probabilities must sum to 1, so need estimate only n-1 parameters...

# Estimating Parameters: $Y$ discrete, $X_i$ continuous

Maximum likelihood estimates:

$$\hat{\mu}_{ik} = \frac{1}{\sum_j \delta(Y^j = y_k)} \sum_j X_i^j \delta(Y^j = y_k)$$

jth training example

ith feature

kth class

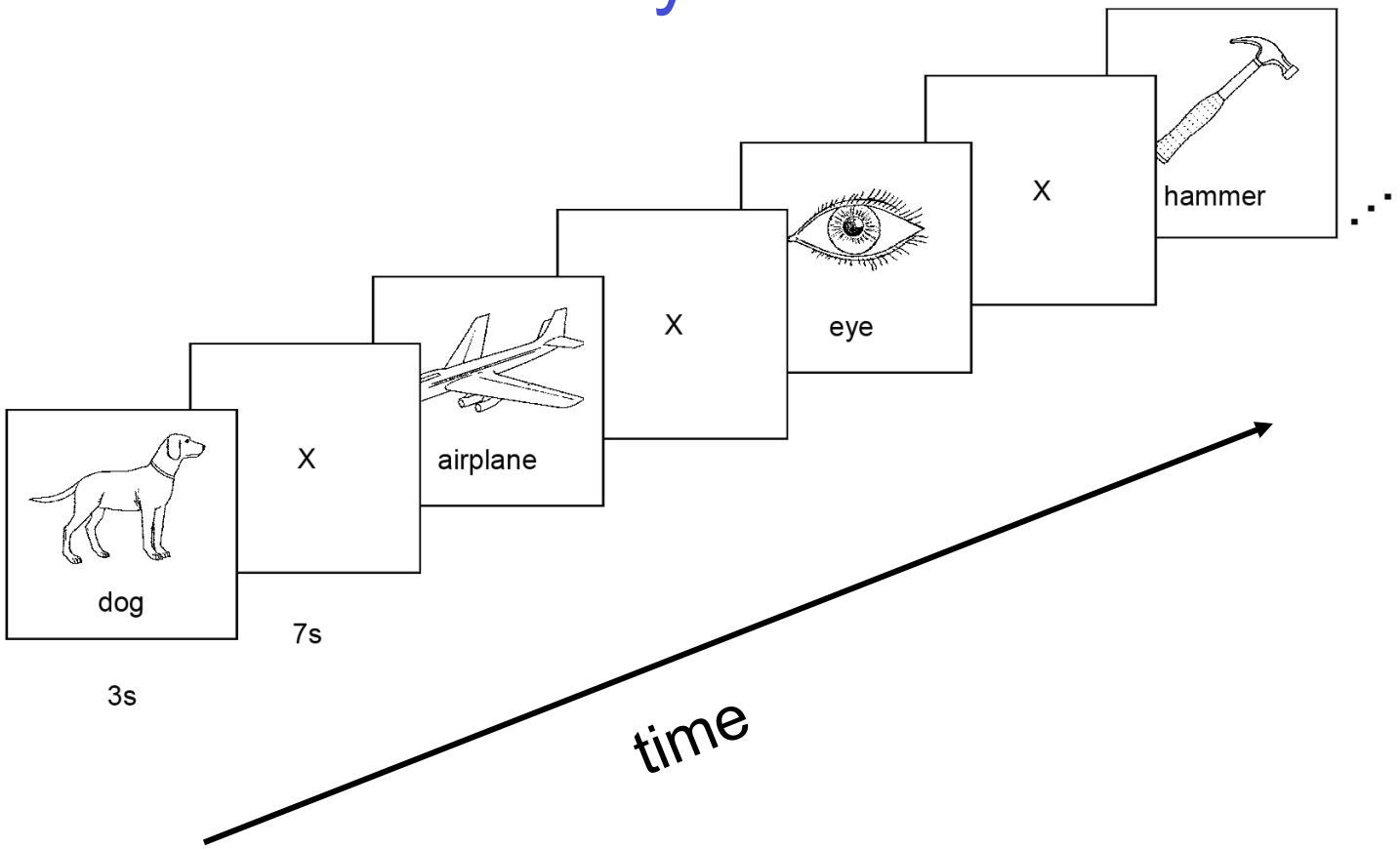
δ(z)=1 if z true, else 0

$$\hat{\sigma}_{ik}^2 = \frac{1}{\sum_j \delta(Y^j = y_k)} \sum_j (X_i^j - \hat{\mu}_{ik})^2 \delta(Y^j = y_k)$$

# GNB Example: Classify a person's cognitive activity, based on brain image

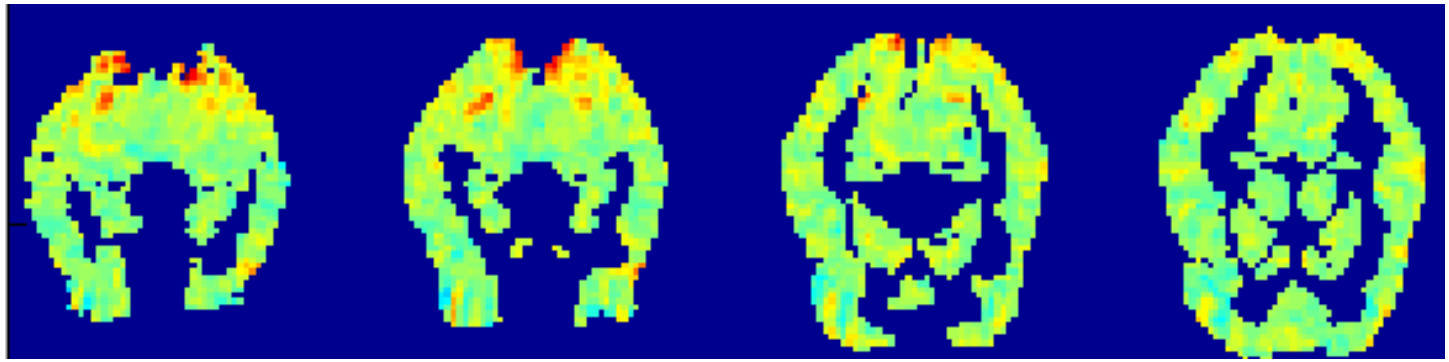
- are they reading a sentence or viewing a picture?
- reading the word "Hammer" or "Apartment"
- viewing a vertical or horizontal line?
- answering the question, or getting confused?

# Stimuli for our study:

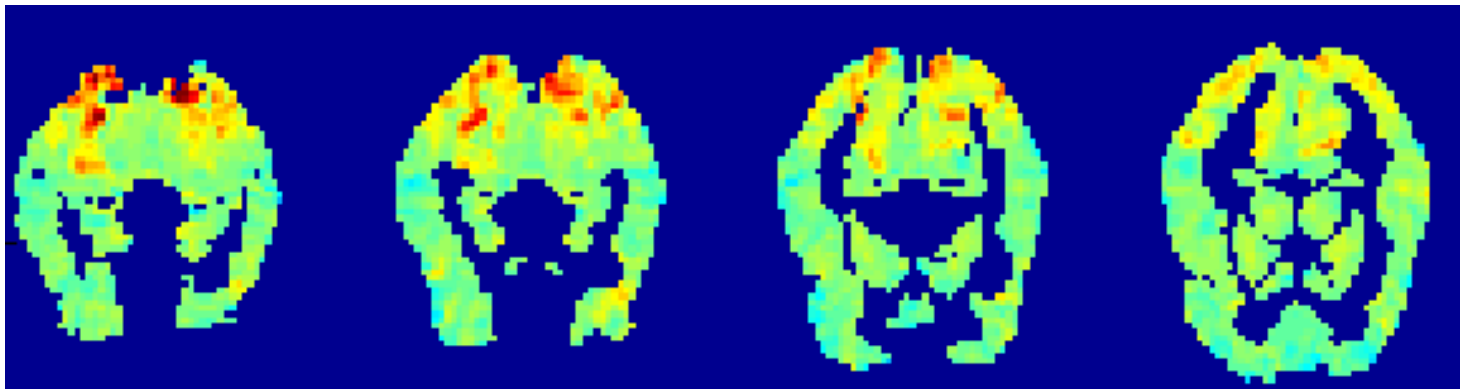



60 distinct exemplars, presented 6 times each

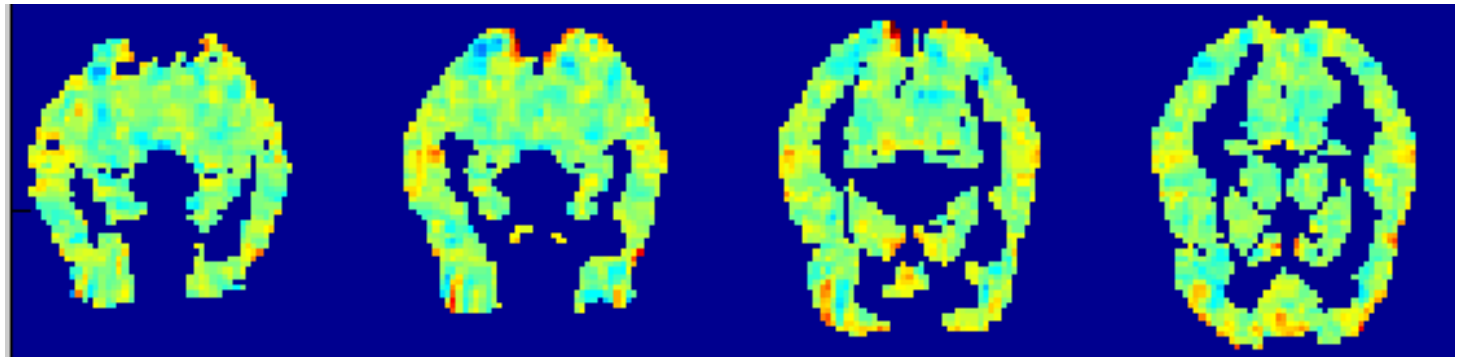
## fMRI voxel means for “bottle”: means defining P(Xi | Y=“bottle)

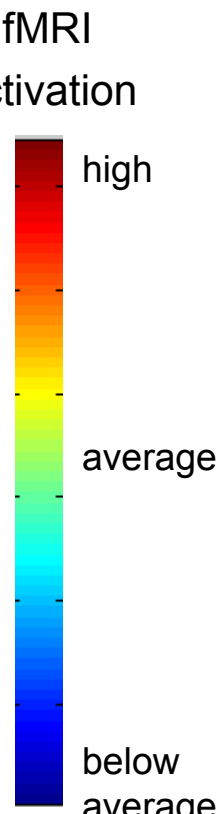


## Mean fMRI activation over all stimuli:

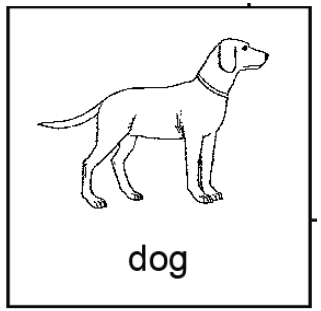


## “bottle” minus mean activation:



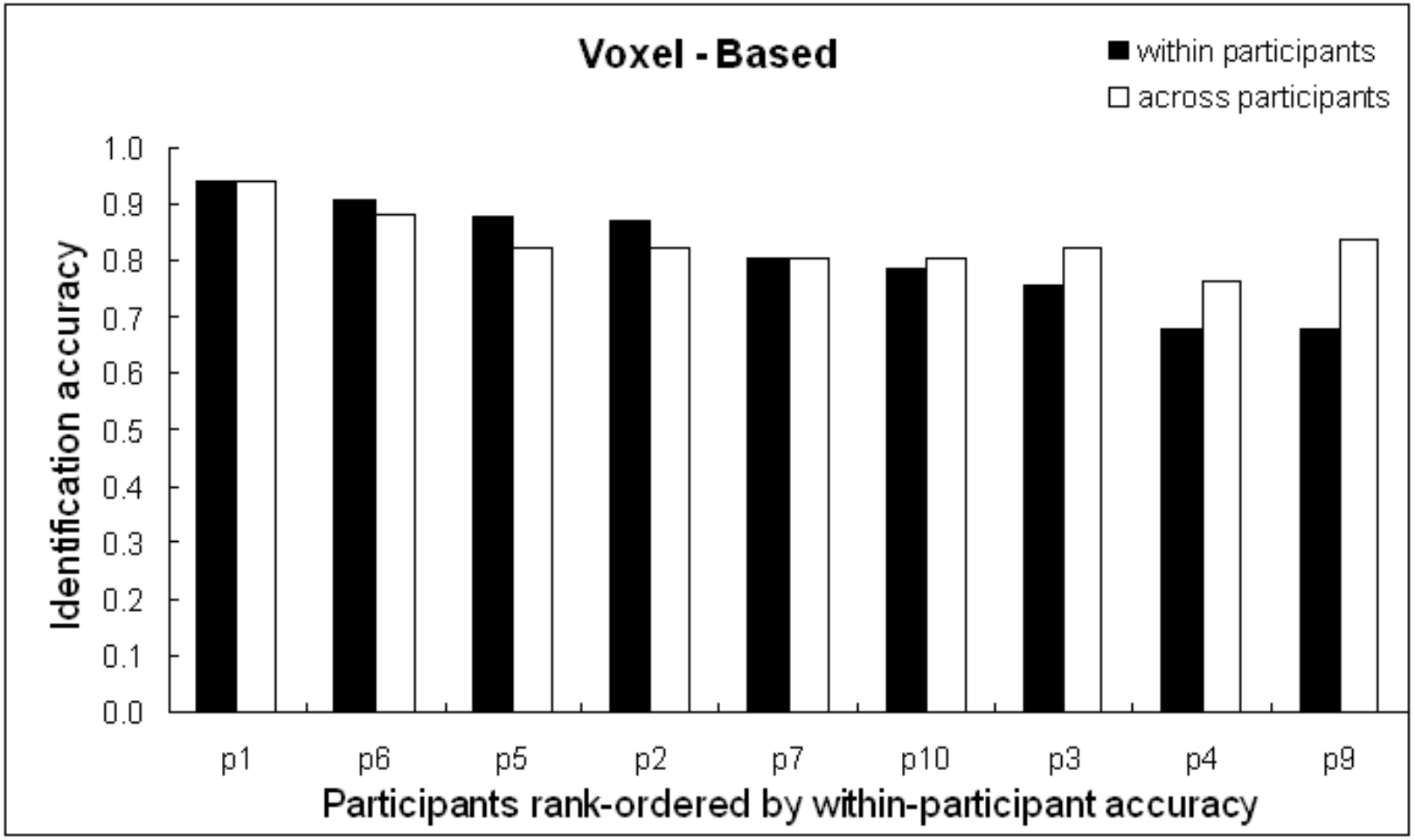


# Scaling up: 60 exemplars



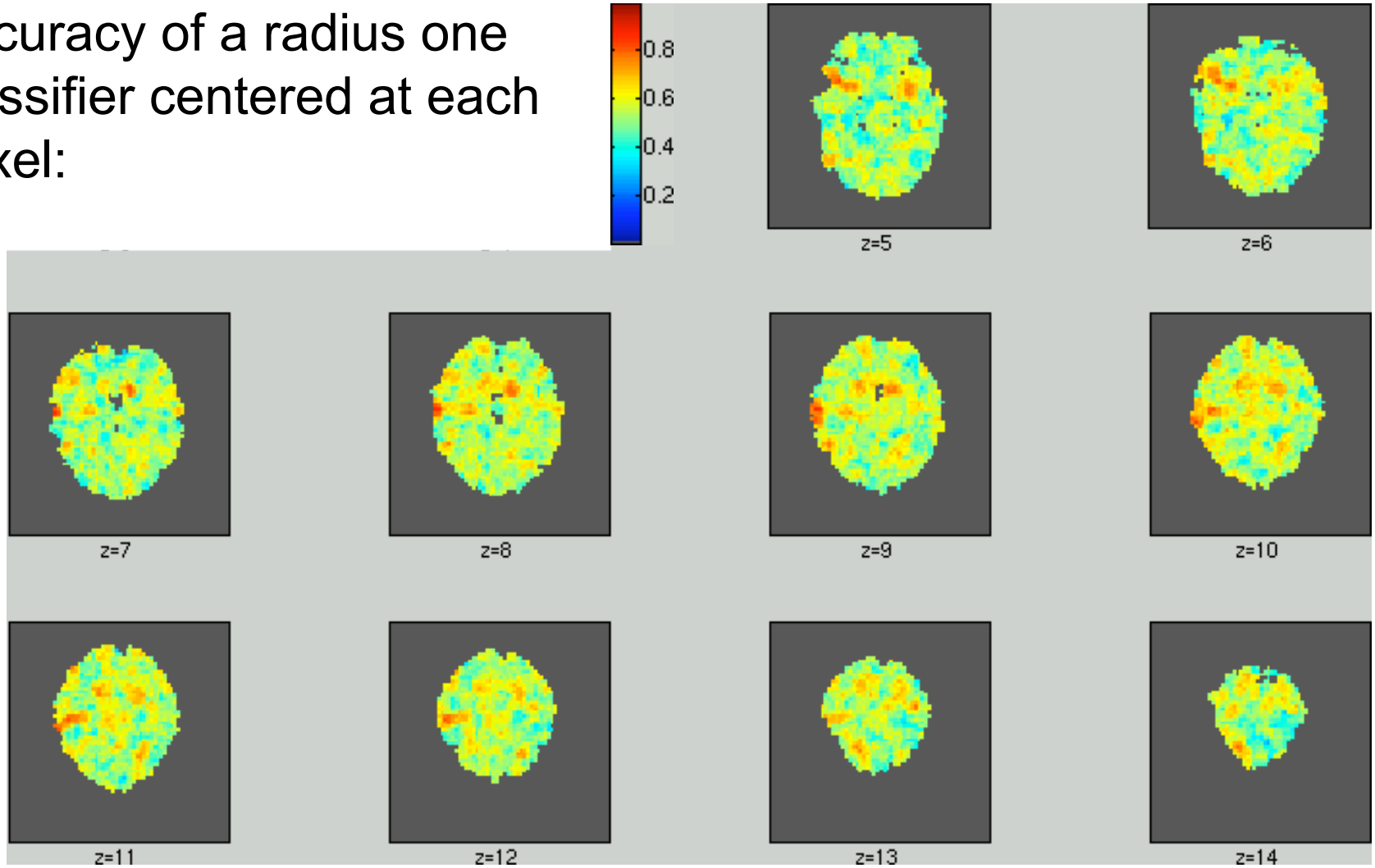

| Categories | Exemplars | | | | |
|---|---|---|---|---|---|
| BODY PARTS | leg | arm | eye | foot | hand |
| FURNITURE | chair | table | bed | desk | dresser |
| VEHICLES | car | airplane | train | truck | bicycle |
| ANIMALS | horse | dog | bear | cow | cat |
| KITCHEN UTENSILS | glass | knife | bottle | cup | spoon |
| TOOLS | chisel | hammer | screwdriver | pliers | saw |
| BUILDINGS | apartment | barn | house | church | igloo |
| PART OF A BUILDING | window | door | chimney | closet | arch |
| CLOTHING | coat | dress | shirt | skirt | pants |
| INSECTS | fly | ant | bee | butterfly | beetle |
| VEGETABLES | lettuce | tomato | carrot | corn | celery |
| MAN MADE OBJECTS | refrigerator | key | telephone | watch | bell |

# Rank Accuracy Distinguishing among 60 words



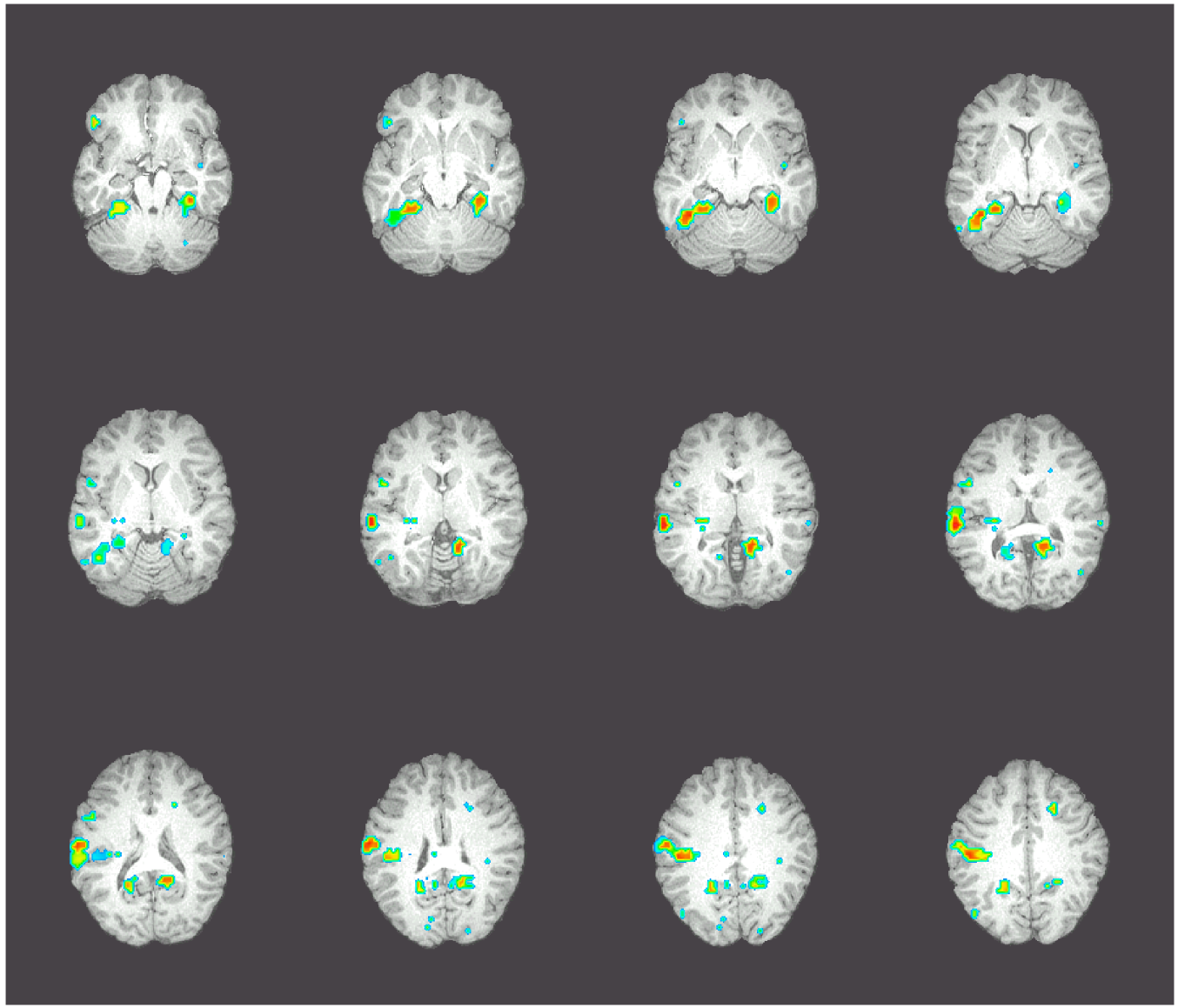
# Where in the brain is activity that distinguishes tools vs. buildings?

Accuracy of a radius one classifier centered at each voxel:



# voxel clusters: searchlights

Accuracies of cubical 27-voxel classifiers centered at each significant voxel [0.7-0.8]



# What you should know:

- Training and using classifiers based on Bayes rule
- Conditional independence
  - What it is
  - Why it's important
- Naïve Bayes
  - What it is
  - Why we use it so much
  - Training using MLE, MAP estimates
  - Discrete variables (Bernoulli) and continuous (Gaussian)

# Questions:

- Can you use Naïve Bayes for a combination of discrete and real-valued $X_i$?
- How can we easily model just 2 of n attributes as dependent?
- What does the decision surface of a Naïve Bayes classifier look like?

# What is form of decision surface for Naïve Bayes classifier?