Bayesian Classifiers, Conditional Independence and Naïve Bayes

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

● Suppose Y=Wealth, X=<Gender, HoursWorked>

Gender	HrsWorked	P(rich | G,HW)	P(poor | G,HW)
F	<40.5	.09	.91
F	>40.5	.21	.79
M	<40.5	.23	.77
M	>40.5	.38	.62

How many parameters must we estimate?

Gender	HrsWorked	P(rich | G,HW)	P(poor | G,HW)
F	<40.5	.09	.91
F	>40.5	.21	.79
M	<40.5	.23	.77
M	>40.5	.38	.62

- Suppose $X = <X_1, \ldots X_n>$

  where X_i and Y are boolean RV's

  To estimate $P(Y | X_1, X_2, \ldots X_n)$

- If we have 30 Xi's instead of 2?

Can we reduce params by using Bayes Rule?

- Suppose $X = <X_1, \ldots X_n>$
- where X_i and Y are boolean RV's

$$P(Y|X) = \frac{\overset{2^n-1}{P(X|Y)} \overset{2}{P(Y)}}{P(X)}$$

Recall 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

朴素贝叶斯假设输入各维度独立 使参数量减少

- 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≠j

独立可以使参数数量变小

$$P(X_1|Y) \, P(X_2|Y) \cdots P(X_n|Y) \, P(Y)$$
$$2n - 1 + 2 = 2n + 1$$

Conditional Independence

Definition: X is <u>conditionally independent</u> 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 Xi are conditionally independent, given Y

- Given this assumption, then:

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

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

- How many parameters to describe $P(X_1...X_n|Y)$? $P(Y)$?
  - Without conditional independent assumption?
  - With conditional independent assumption?

Naïve Bayes in a Nutshell

Bayes rule:

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

Assuming conditional independence among X_i's:

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

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

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

Naïve Bayes Algorithm – discrete Xi

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

  x_{ij}：第 i 个特征 X_i 的第 j 个可能取值

  X_i $P(x_{ij} | Y = y_k)$

  x_{i1} θ_{i1k}

  $\vdots$ $\vdots$

  x_{ij} θ_{ijk}

  $\vdots$ $\vdots$

  x_{iL} θ_{iLk}

- 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):

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

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

似然函数

最大似然的缺陷:
没有考虑参数的先验
(这是指θijk的先验)

Number of items in
dataset D for which Y=y$_k$

Naïve Bayes: Subtlety #1

- If unlucky, our MLE estimate for $P(X_i | Y)$ 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?

Estimating Parameters

- Maximum Likelihood Estimate (MLE): choose θ that maximizes probability of observed data $\mathcal{D}$

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

- Maximum a Posteriori (MAP) estimate: choose θ that is most probable given prior probability and the data

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

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

P(D|θ)和P(θ)等形成某函数：多类 多顶式一利利贵雷

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 (Beta, Dirichlet priors):**

$$\hat{\pi}_k = \hat{P}(Y = y_k) = \frac{\#D\{Y = y_k\} + \alpha_k}{|D| + \sum_m \alpha_m}$$
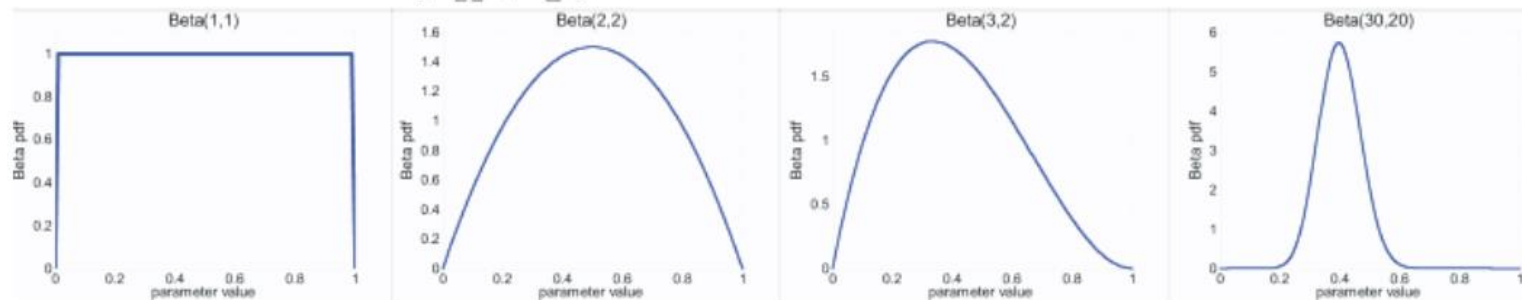
Only difference: "imaginary" examples

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

Beta prior distribution – P(θ)

$$P(\theta) = \frac{\theta^{\beta_H - 1}(1 - \theta)^{\beta_T - 1}}{B(\beta_H, \beta_T)} \sim Beta(\beta_H, \beta_T)$$
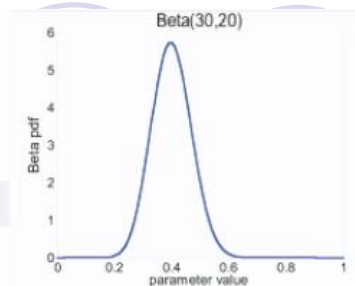
Mean:

Mode:

- Likelihood function: $P(\mathcal{D} \mid \theta) = \theta^{\alpha_H}(1 - \theta)^{\alpha_T}$
- Posterior: $P(\theta \mid \mathcal{D}) \propto P(\mathcal{D} \mid \theta)P(\theta)$

MAP for Beta distribution

Beta(30,20)

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

■ MAP: use most likely parameter:

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

- ■ Beta prior equivalent to extra thumbtack flips
- ■ As $N \to \infty$, prior is "forgotten"
- ■ **But, for small sample size, prior is important!**

Dirichlet distribution

- **number of heads in N flips of a two-sided coin**
  - follows a binomial distribution
  - Beta is a good prior (conjugate prior for binomial)

- **what it's not two-sided, but k-sided?**
  - follows a multinomial distribution
  - Dirichlet distribution is the conjugate prior

$$P(\theta_1, \theta_2, ...\theta_K) = \frac{1}{B(\alpha)} \prod_i^K \theta_i^{(\alpha_1 - 1)}$$

Lejeune Dirichlet

Johann Peter Gustav Lejeune Dirichlet

Born	13 February 1805 Düren, French Empire
Died	5 May 1859 (aged 54) Göttingen, Hanover
Residence	Germany
Nationality	German
Fields	Mathematician
Institutions	University of Berlin University of Breslau University of Göttingen
Alma mater	University of Bonn
Doctoral advisor	Simeon Poisson Joseph Fourier
Doctoral students	Ferdinand Eisenstein Leopold Kronecker Rudolf Lipschitz Carl Wilhelm Borchardt
Known for	Dirichlet function Dirichlet eta function

Naïve Bayes: Subtlety #2

- Often the X_i are not really conditionally independent 不独立： $P(X_1,\ldots,X_n|Y) \neq \prod_{i=1}^{n} P(X_i|Y)$

- 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: $X_i = X_k$

$P(X_1,\ldots,X_n|Y=1)$
$= P(X_1|Y=1)\cdots P(X_j,X_k|Y=1)\cdots P(X_n|Y=1)$
$\xrightarrow{\alpha^2} P(X_j|Y=1)\,P(X_k|Y=1)$
老师没有考虑独立

$P(X_1,\ldots,X_n|Y=2)$
$= P(X_1|Y=2)\cdots P(X_j,X_k|Y=2)\cdots P(X_n|Y=2)$
$\xrightarrow{\beta^2} P(X_j|Y=2)\,P(X_k|Y=2)$

过拟合
可能导致决策的偏倒
极而分析

$$P(y|x) = \frac{P(x_1, x_2, \ldots, x_n|y)P(y)}{P(x)}$$

若假设没独立 $P(x_1, x_2, \ldots, x_n|y=1)$

$$= P(x_1|y=1) \cdots \overset{\alpha}{P(x_i|y=1)} \cdots \overset{\alpha}{P(x_j|y=1)} \cdots P(x_n|y=1) \qquad \alpha^2 P_a$$

$P(x_1, x_2, \ldots, x_n|y=2)$

$$= P(x_1|y=2) \cdots \overset{\beta}{P(x_i|y=2)} \cdots \overset{\beta}{P(x_j|y=2)} \cdots P(x_n|y=2) \qquad \beta^2 P_b$$

事实上, $P(x_1, x_2, \ldots, x_n|y=1)$

$$= P(x_1|y=1) \cdots P(x_n|y=1) \, \overset{\alpha}{P(x_i, x_j|y=1)} \qquad \alpha P_a$$

$P(x_1, x_2, \ldots, x_n|y=2)$

$$= P(x_1|y=2) \cdots P(x_n|y=2) \, \overset{\beta}{P(x_i, x_j|y=2)} \qquad \beta P_b$$

若 $\alpha P_a > \beta P_b$ 可能存在 α, β 使 $\alpha^2 P_a < \beta^2 P_b$ 导致决策颠倒

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?

Baseline: Bag of Words Approach

Learning to classify text

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

1. Represent each document by vector of words   *eg. 垃圾邮件分类*
   - one attribute per word position in document

2. Learning: Use training examples to estimate
   - $P(+)$ *用文本中的每个词表示其属性*
   - $P(-)$
   - $P(doc|+)$ *词频*
   - $P(doc|-)$

Naive Bayes conditional independence assumption

拆分成文章中的词

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

朴素bayes可以不等长；不涉及其他文本
KNN要求等长；涉及其他文本

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

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

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	

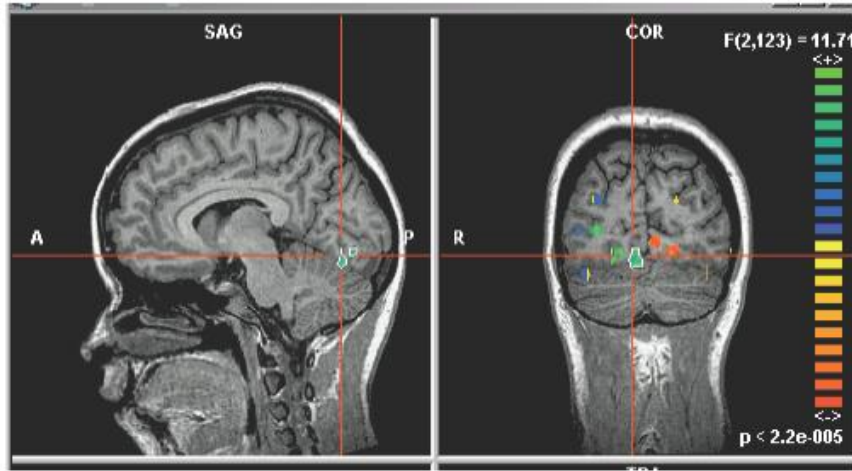Naive Bayes: 89% classification accuracy

Learning curve for 20 newsgroups
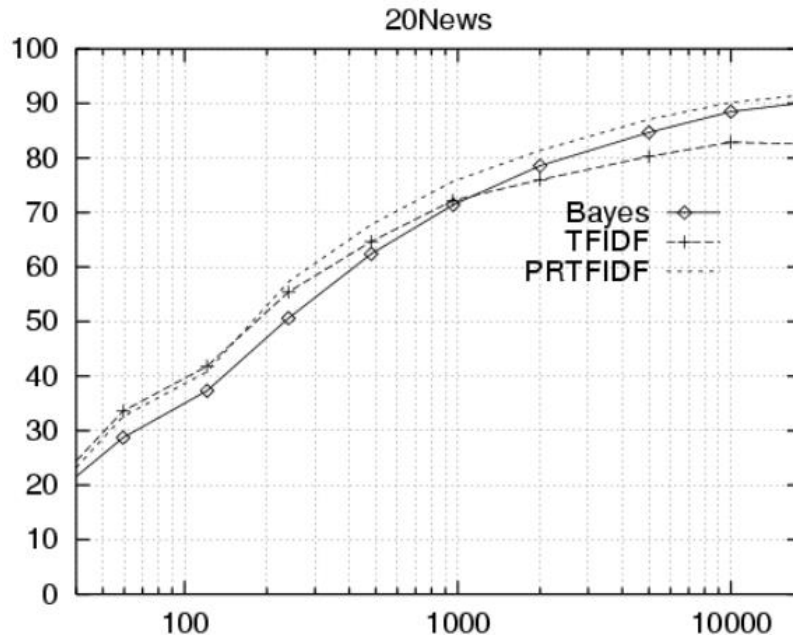
Accuracy vs. Training set size (1/3 withheld for test)

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

Still have:
$$P(Y = y_k | X_1 \ldots X_n) = \frac{P(Y = y_k) \prod_i P(X_i | Y = y_k)}{\sum_j P(Y = y_j) \prod_i P(X_i | Y = y_j)}$$

Just need to decide how to represent P(Xi | Y)

连续情况：每个维度都在连续空间中取值 $X_i \in \mathbb{R}$

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   其它情况二次项无法抵消
  - is independent of Y (i.e., σ_i), 各类别方差一样：线性分类器
  - or independent of X_i (i.e., σ_k) 各维方差一样
  - or both (i.e., σ) 所有方差都一样

伯努利分布 or 多项式分布

$P(Y=y_k|X) \propto P(X=x_1|Y=y_k)\cdots P(X=x_m|Y=y_k)P(Y=y_k)$

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 μ_{ik}, variance σ_{ik}

- Classify (X^{new})

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

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

Estimating Parameters: *Y* discrete, *X*$_i$ continuous

样本→总体
最大似然

Maximum likelihood estimates:

jth training example

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

ith feature

kth class

δ(z)=1 if z true, else 0

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

（满足 y=yk 的所有 Xi 的均值和方差）

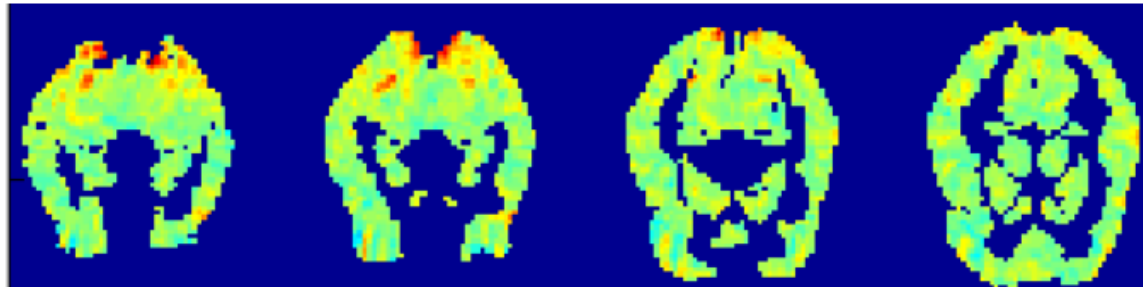# GNB Example: Classify a person's cognitive activity, based on brain image

- 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:

dog
3s
7s
airplane
X
eye
X
hammer
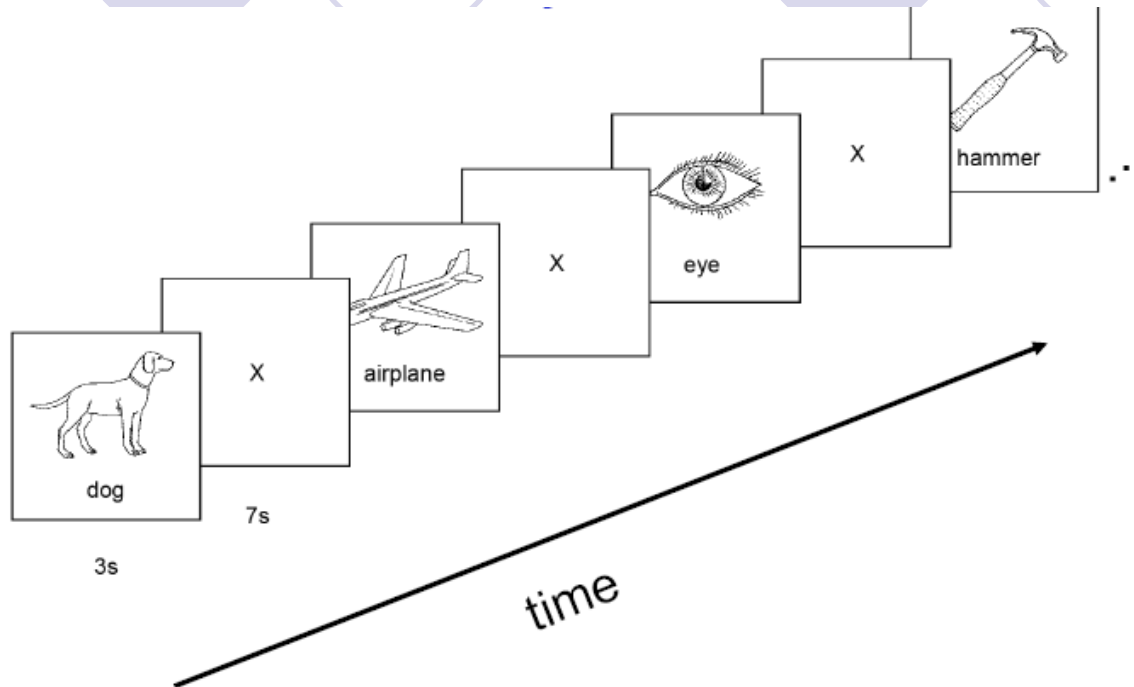
time

60 distinct exemplars, presented 6 times each

fMRI voxel means for "bottle": means defining P(Xi | Y="bottle)

Mean fMRI activation over all stimuli:
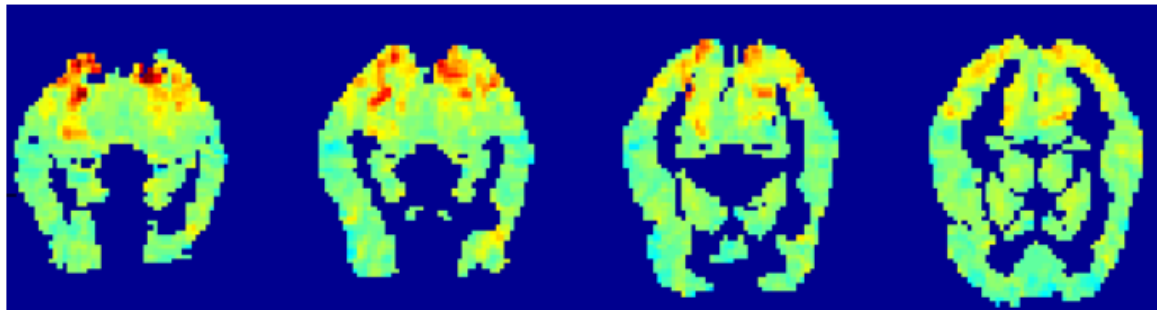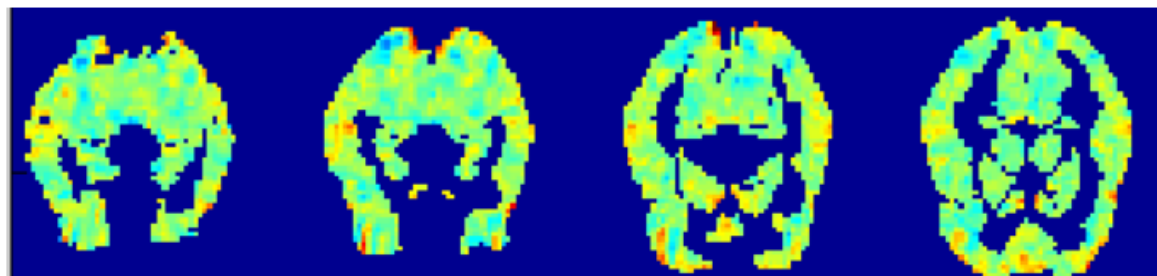

"bottle" minus mean activation:

fMRI activation

high

average

below average

Rank Accuracy Distinguishing among 60 words
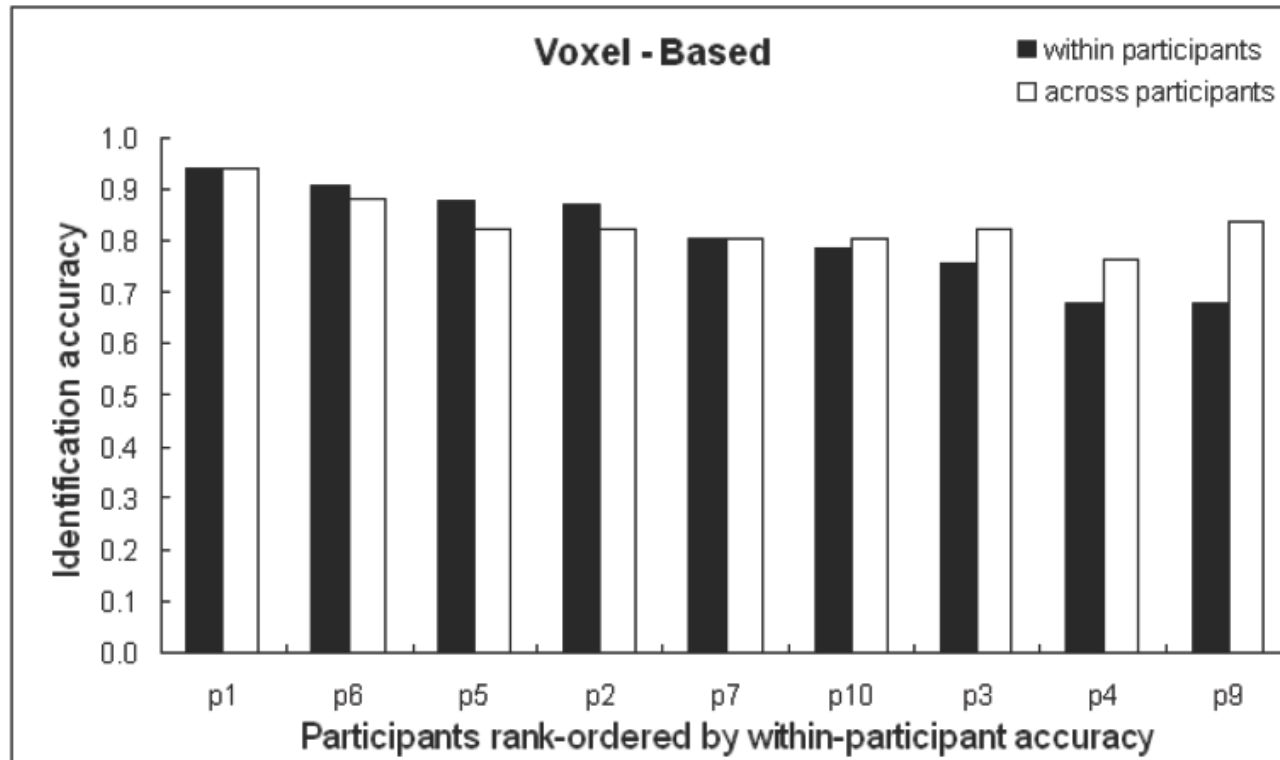
What you should know:

- **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 and continuous (Gaussian)

Questions:

- What is the error will classifier achieve if Naïve Bayes assumption is satisfied and we have infinite training data?

  Bayes分类器最小错误率

- Can you use Naïve Bayes for a combination of discrete and real-valued X_i?

  一些维度连续，一些维度离散

- How can we easily model just 2 of n attributes as dependent?

  变成联合概率，若 x_i, x_j 不符合条件独立: $\frac{p(x_i|y_k)}{p(x_j|y_k)} \times \Rightarrow p(x_i, x_j|y_k)$

- What does the decision surface of a Naïve Bayes classifier look like?

  不一定是线性的

  离散
  连续: x_i的方差独立于 y 时是线性