Classification - Naïve Bayes with MLE/MAP

Aarti Singh

Machine Learning 10-315
Sept 9, 2019

Naïve Bayes Classifier

- Bayes Classifier with additional "naïve" assumption:
  - Features are independent given class:

$$P(X_1...X_d|Y) = \prod_{i=1}^{d} P(X_i|Y)$$

$$f_{NB}(\mathbf{x}) = \arg\max_{y} \; P(x_1,\ldots,x_d \mid y)P(y)$$

$$= \arg\max_{y} \prod_{i=1}^{d} P(x_i|y)P(y)$$

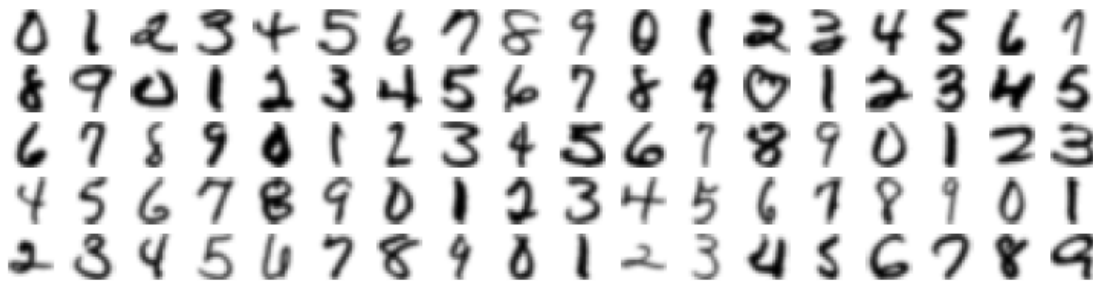- Has fewer parameters, and hence requires fewer training data

Linear instead of Quadratic or Exponential in d!

Hand-written digit recognition

Input, X (images of hand-written digits) ⮕ Label, Y

0, 1, 2, ..., 9

Black-white images – d-dim binary (0/1) vector

Discrete Naïve Bayes model:

$P(Y = y) = p_y$ for all y in 0, 1, ..., 9 $\qquad$ $p_0, p_1, ..., p_9$ (sum to 1)

Akin to roll of a dice (Multinomial)

$P(X_i = x_i | Y = y)$ - one probability value for each y, pixel i

Akin to a coin flip for each y and pixel i (Bernoulli)

3

MLE for Bernoulli, Multinomial

Bernoulli(θ)

P(Heads) = θ, P(Tails) = 1-θ

$$\widehat{\theta}_{MLE} \;=\; \frac{\alpha_H}{\alpha_H + \alpha_T}$$ "Frequency of heads"

Multinomial(θ={p_1,p_2,..,p_6})

P(1) = p_1, P(2) = p_2, …, P(6) = p_6 = 1-(p_1+…p_5)

$$\widehat{\theta}_{MLE} \;=\; \hat{p}_{1,MLE}, \dots, \hat{p}_{6,MLE}$$

$$\hat{p}_{y,MLE} = \frac{\alpha_y}{\sum_y \alpha_y}$$ "Frequency of roll y"

Naïve Bayes Algo – Discrete features

- Training Data $\{(X^{(j)}, Y^{(j)})\}_{j=1}^n$ $X^{(j)} = (X_1^{(j)}, \ldots, X_d^{(j)})$

- Maximum Likelihood Estimates

  – For Class probability

$$\widehat{P}(y) = \frac{\{\#j : Y^{(j)} = y\}}{n}$$

  – For class conditional distribution

$$\hat{P}(x_i|y) = \frac{\widehat{P}(x_i, y)}{\widehat{P}(y)} = \frac{\{\#j : X_i^{(j)} = x_i, Y^{(j)} = y\}/n}{\{\#j : Y^{(j)} = y\}/n}$$

- NB Prediction for test data $X = (x_1, \ldots, x_d)$

$$Y = \arg\max_y \widehat{P}(y) \prod_{i=1}^d \frac{\widehat{P}(x_i, y)}{\widehat{P}(y)}$$

Issues with Naïve Bayes

- **Issue 1:** Usually, features are not conditionally independent:

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

Nonetheless, NB is the single most used classifier particularly when data is limited, works well

- **Issue 2:** Typically use MAP estimates instead of MLE since insufficient data may cause MLE to be zero.

Insufficient data for MLE

- What if you never see a training instance where X_1=a when Y=b?

    - e.g., b={SpamEmail}, a ={'Earn'}
    - P(X_1= a | Y = b) = 0

- Thus, no matter what the values X_2,…,X_d take:

$$\boxed{\widehat{P}(X_1 = a, X_2...X_n|Y) = \widehat{P}(X_1 = a|Y) \prod_{i=2}^{d} \widehat{P}(X_i|Y)} = 0$$

- What now???

Naïve Bayes Algo – Discrete features

- Training Data $\{(X^{(j)}, Y^{(j)})\}_{j=1}^n$ $X^{(j)} = (X_1^{(j)}, \ldots, X_d^{(j)})$

- Maximum A Posteriori (MAP) Estimates – add m "virtual" datapts
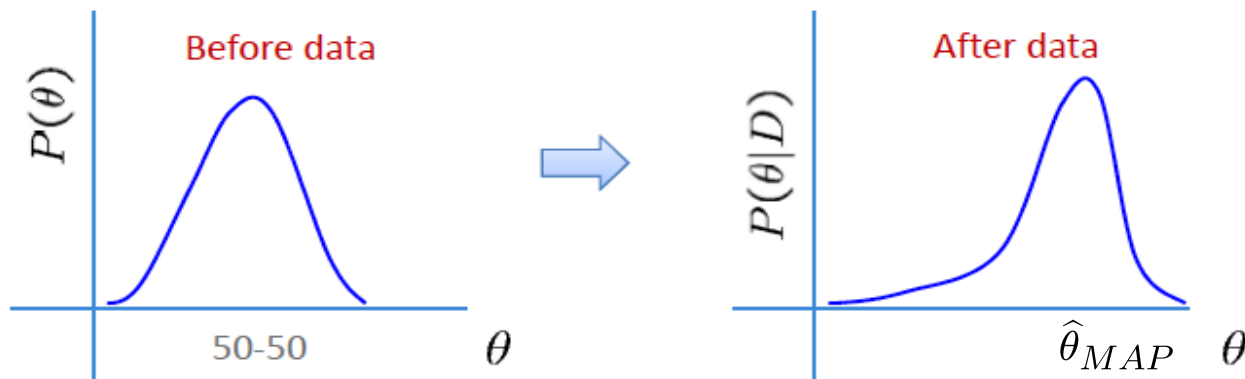
  Assume priors

  $$Q(Y = b) \qquad\qquad Q(X_i = a, Y = b)$$

$$\widehat{P}(X_i = a | Y = b) = \frac{\{\#j : X_i^{(j)} = a, Y^{(j)} = b\} + mQ(X_i = a, Y = b)}{\{\#j : Y^{(j)} = b\} + \underbrace{mQ(Y = b)}}$$

virtual examples
with Y = b

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

Max A Posteriori (MAP) estimation

Justification for adding virtual examples

- Assume a prior distribution for parameters θ (reflects prior belief in what values θ is likely to take)

- Choose θ that maximizes a posterior probability

$$\widehat{\theta}_{MAP} \;=\; \arg\max_{\theta}\;\; P(\theta \mid D)$$

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

MAP estimation for Bernoulli r.v.

Choose θ that maximizes a posterior probability

$$\widehat{\theta}_{MAP} = \arg\max_{\theta} \quad P(\theta \mid D)$$
$$= \arg\max_{\theta} \quad P(D \mid \theta)P(\theta)$$
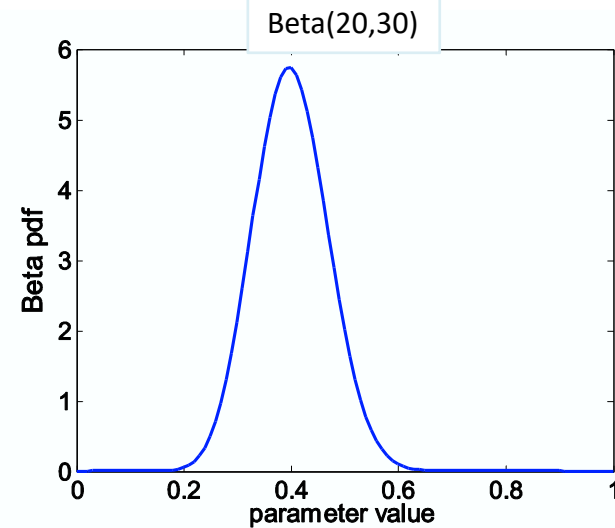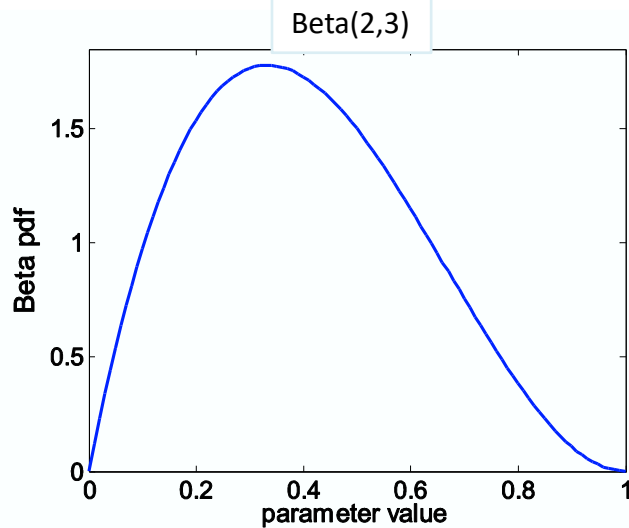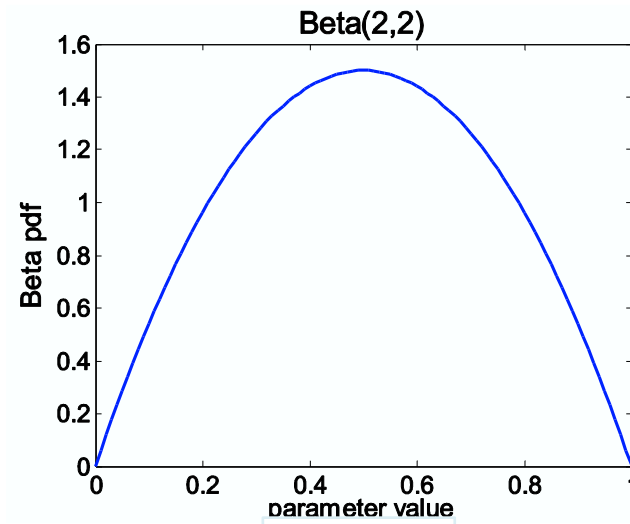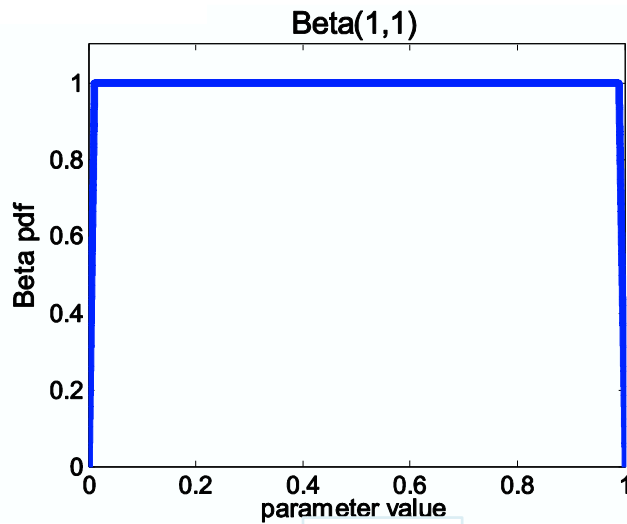
MAP estimate of probability of head:

$$P(\theta) \sim Beta(\beta_H, \beta_T)$$

Beta distribution

$Beta(\beta_H, \beta_T)$ More concentrated as values of β_H, β_T increase

MAP estimation for Bernoulli r.v.

Choose θ that maximizes a posterior probability

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

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

MAP estimate of probability of head:

$$P(\theta) \sim Beta(\beta_H, \beta_T)$$

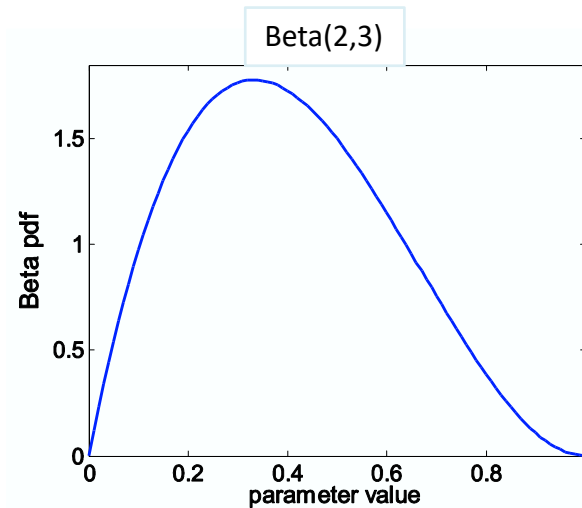Count of H/T simply get added to parameters

$$P(\theta|D) \sim Beta(\beta_H + \alpha_H, \beta_T + \alpha_T)$$

Beta conjugate prior

$$P(\theta) \sim Beta(\beta_H, \beta_T)$$

$$P(\theta|D) \sim Beta(\beta_H + \alpha_H, \beta_T + \alpha_T)$$

Beta(2,2)

Beta(2,3)

After observing 1 Tail

Beta(20,30)

After observing 18 Heads and 28 Tails

As $n = \alpha_H + \alpha_T$ increases, posterior distribution becomes more concentrated

MAP estimation for Bernoulli r.v.

Choose θ that maximizes a posterior probability

$$\widehat{\theta}_{MAP} \; = \; \arg\max_{\theta} \quad P(\theta \mid D)$$

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

MAP estimate of probability of head:

$$P(\theta) \sim Beta(\beta_H, \beta_T)$$

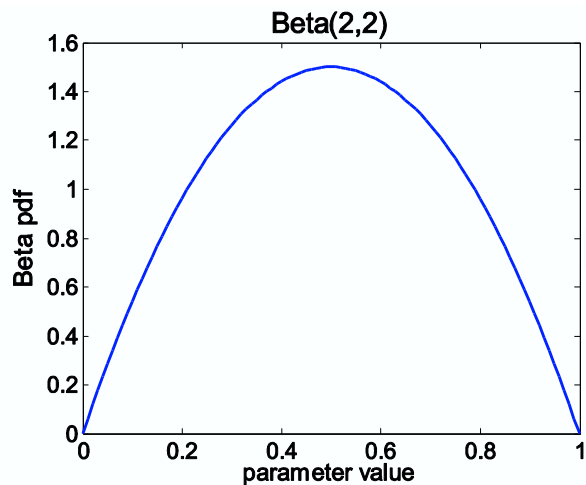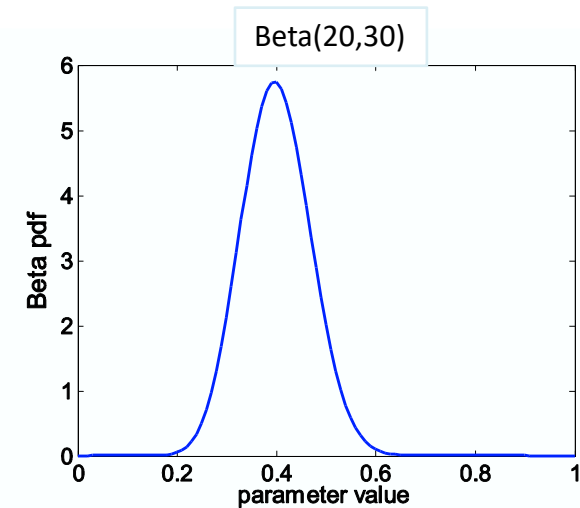Count of H/T simply get added to parameters

$$P(\theta|D) \sim Beta(\beta_H + \alpha_H, \beta_T + \alpha_T)$$

$$\widehat{\theta}_{MAP} \; = \; \frac{\alpha_H + \beta_H - 1}{\alpha_H + \beta_H + \alpha_T + \beta_T - 2}$$

Mode of Beta distribution

Equivalent to adding extra coin flips (β_H - 1 heads, β_T - 1 tails)

As we get more data, effect of prior is "washed out"

MLE vs. MAP

- Maximum Likelihood estimation (MLE)

  Choose value that maximizes the probability of observed data

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

- Maximum *a posteriori* (MAP) estimation

  Choose value that is most probable given observed data and prior belief

  $$\widehat{\theta}_{MAP} = \arg\max_{\theta} P(\theta|D)$$
  $$= \arg\max_{\theta} P(D|\theta)P(\theta)$$

When is MAP same as MLE?
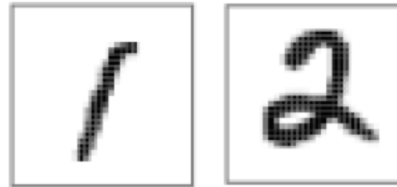
How to learn parameters from data? MLE, MAP

(Continuous case)

Naïve Bayes with continuous features

Training Data:

Each image represented as a vector of **intensity values at the d pixels (features)**

Input, X $\qquad$ ... n greyscale images $\qquad$ $X = \begin{bmatrix} X_1 \\ X_2 \\ \dots \\ X_d \end{bmatrix}$
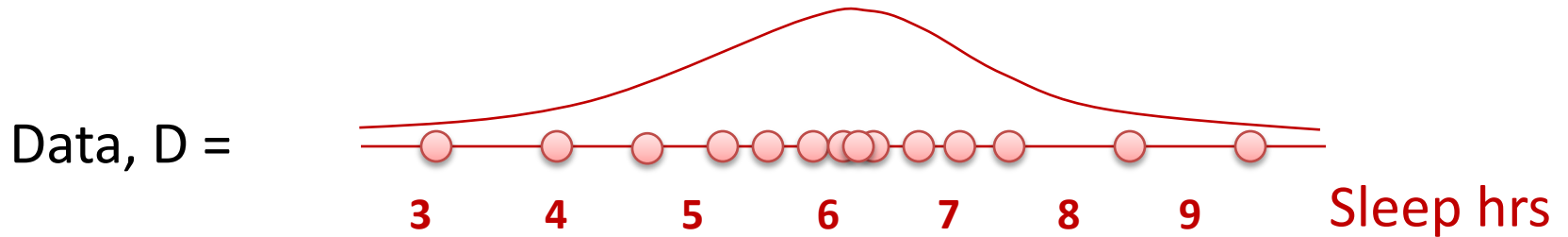
Label, Y $\qquad$ 1 $\qquad$ 2 $\qquad$ ... n labels

Gaussian Naïve Bayes model:

$P(Y = y) = p_y$ for all y in 0, 1, 2, ..., 9 $\qquad$ $p_0, p_1, ..., p_9$ (sum to 1)

$P(X_i = x_i | Y = y) \sim N(\mu^{(y)}_i, \sigma^{2}_i{}^{(y)})$ for each y and each pixel i

Gaussian distribution

Data, D =

3   4   5   6   7   8   9   Sleep hrs

- Parameters: μ – mean, σ^2 - variance

- Sleep hrs are **i.i.d.**:
  - **Independent** events
  - **Identically distributed** according to Gaussian distribution

Maximum Likelihood Estimation (MLE)

Choose $\theta = (\mu, \sigma^2)$ that maximizes the probability of observed data

$$\hat{\theta}_{MLE} = \arg\max_\theta P(D \mid \theta)$$

$$= \arg\max_\theta \prod_{i=1}^{n} P(X_i \mid \theta) \qquad \text{Independent draws}$$

$$= \arg\max_\theta \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(X_i - \mu)^2 / 2\sigma^2} \qquad \text{Identically distributed}$$

$$= \arg\max_{\theta = (\mu, \sigma^2)} \underbrace{\frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\sum_{i=1}^{n}(X_i - \mu)^2 / 2\sigma^2}}_{J(\theta)}$$

MLE for Gaussian mean and variance

$$\widehat{\mu}_{MLE} \;=\; \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\widehat{\sigma}^2_{MLE} \;=\; \frac{1}{n}\sum_{i=1}^{n} (x_i - \widehat{\mu})^2$$

<u>Note:</u> MLE for the variance of a Gaussian is **biased**

- Expected result of estimation is **not** true parameter!
- Unbiased variance estimator:

$$\widehat{\sigma}^2_{unbiased} \;=\; \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \widehat{\mu})^2$$

Naïve Bayes Algo – continuous features

- Training Data $\{(X^{(j)}, Y^{(j)})\}_{j=1}^{n}$ $\qquad$ $X^{(j)} = (X_1^{(j)}, \ldots, X_d^{(j)})$

- Maximum Likelihood Estimates
  - For Class probability
  $$\hat{P}(y) = \frac{\{\#j : Y^{(j)} = y\}}{n}$$
  - For class conditional distribution
  $$\hat{P}(x_i | y) = N(\hat{\mu}_i^{(y)}, \hat{\sigma}_i^{2(y)})$$

  MLE estimates

- NB Prediction for test data $\qquad X = (x_1, \ldots, x_d)$
$$Y = \arg\max_y \hat{P}(y) \prod_{i=1}^{d} \hat{P}(x_i | y)$$

Naïve Bayes Algo – continuous features

Maximum likelihood estimates:

$$\widehat{\mu}_{MLE} = \frac{1}{n}\sum_{j=1}^{n} x_j$$

$$\hat{\mu}_i{}^{(y)} = \frac{1}{\sum_j 1_{(Y^{(j)}=y)}} \sum_j X_i^{(j)} 1_{(Y^{(j)}=y)}$$

ith pixel in jth training image

y class

jth training image

$$\hat{\sigma}^2_{unbiased} = \frac{1}{n-1}\sum_{j=1}^{n}(x_j - \hat{\mu}_{\mathrm{MLE}})^2$$

$$\hat{\sigma_i}^2(y) = \frac{1}{\sum_j 1_{(Y^{(j)}=y)} - 1} \sum_j (X_i^{(j)} - \hat{\mu}_i^{(y)})^2 1_{(Y^{(j)}=y)}$$

MAP estimation for Gaussian r.v.

Parameters $\theta = (\mu, \sigma^2)$

- Mean μ: Gaussian prior $= N(\eta, \lambda^2)$

$$P(\mu \mid \eta, \lambda) = \frac{1}{\lambda\sqrt{2\pi}} e^{\frac{-(\mu-\eta)^2}{2\lambda^2}}$$

$$\widehat{\mu}_{MAP} = \frac{\frac{1}{\sigma^2}\sum_{i=1}^{n} x_i + \frac{\eta}{\lambda^2}}{\frac{n}{\sigma^2} + \frac{1}{\lambda^2}} \qquad \widehat{\mu}_{MLE} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

As we get more data, effect of prior is "washed out"

- Variance σ^2: Inverse-Wishart Distribution

Applications of Naïve Bayes

- Text classification

    Documents ➜ Topic

    Emails ➜ Spam

    Bag of words example discussed in recitation

- Image classification

    Hand-written characters ➜   Digit, Letter

    Brain scans   ➜   words, objects person is reading/
                                watching

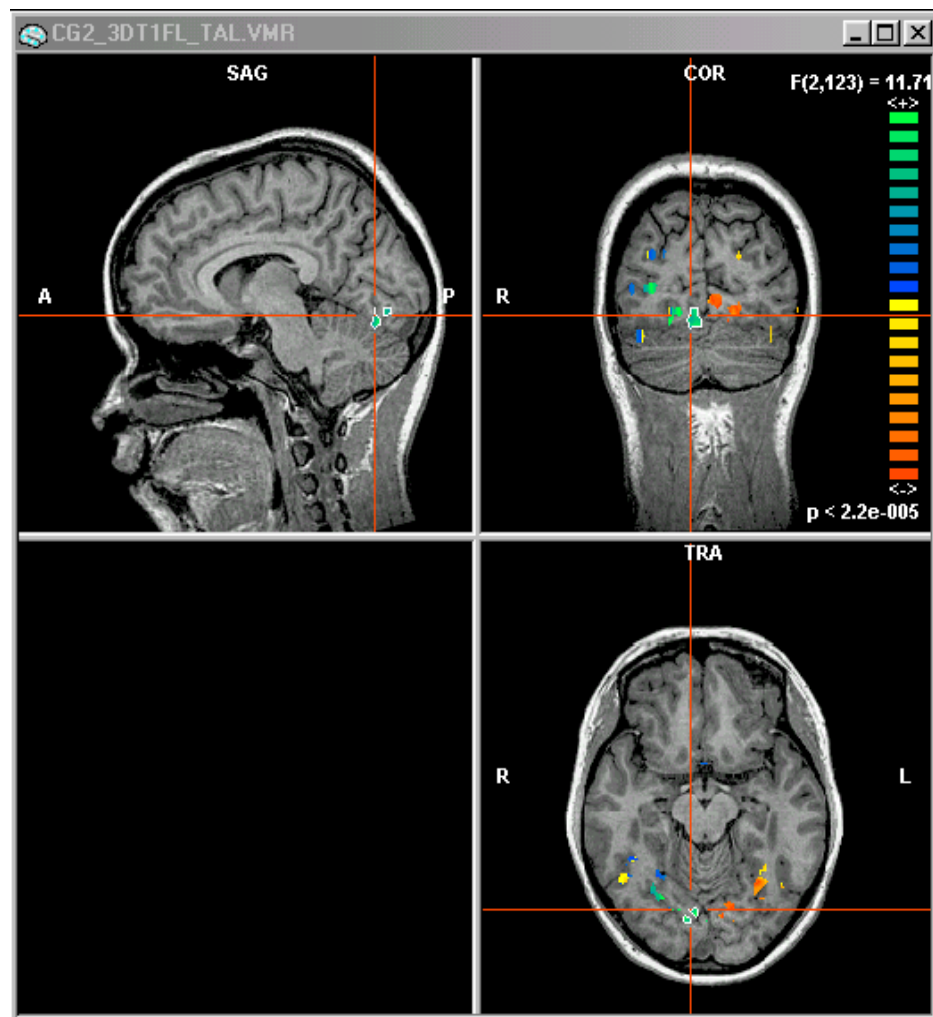# GNB for classifying word category

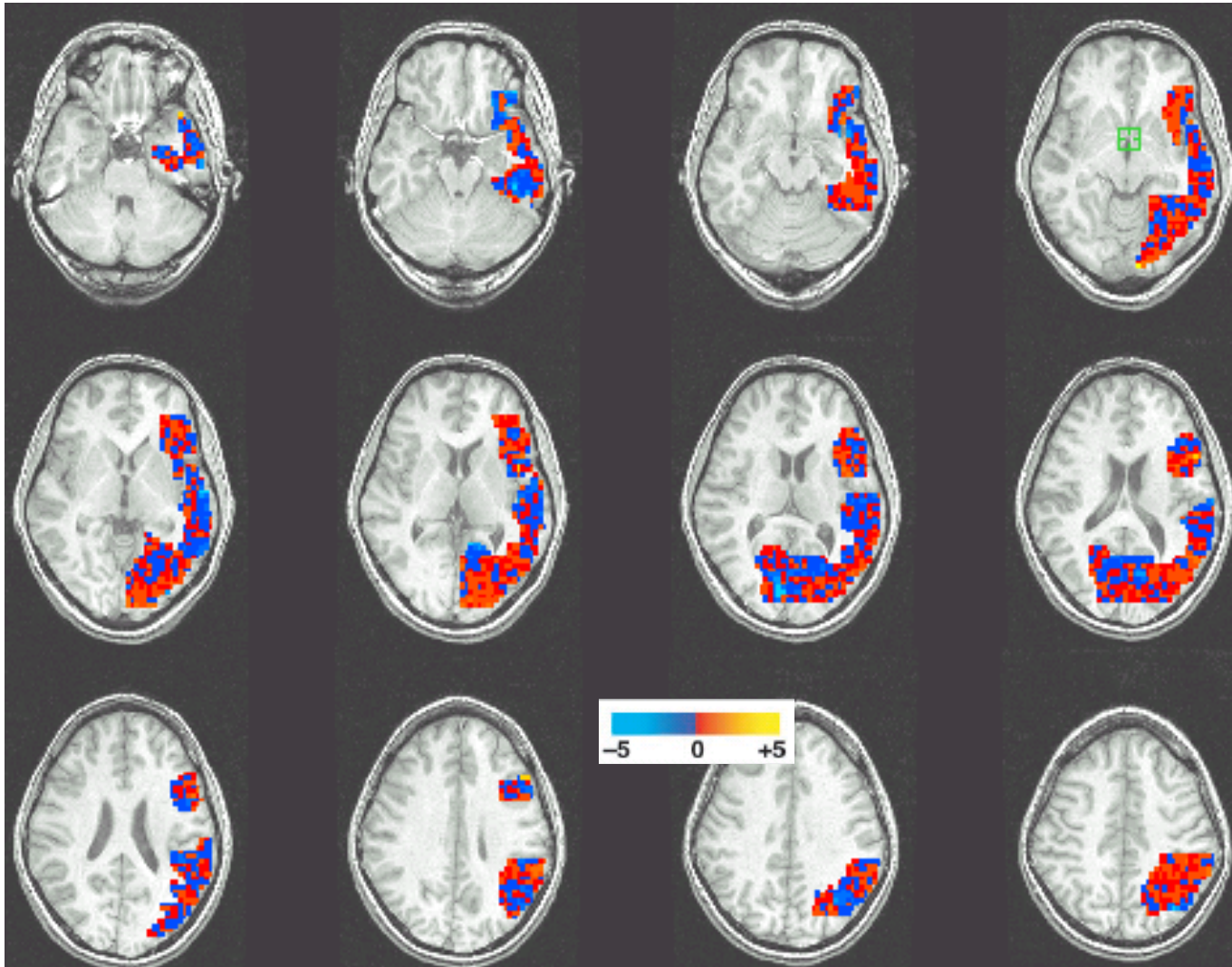[Mitchell et al.03]

~1 mm resolution

~2 images per sec.

15,000 voxels/image

non-invasive, safe

measures Blood Oxygen
Level Dependent (BOLD)
response

Gaussian Naïve Bayes: Learned $\mu_{voxel,word}$

[Mitchell et al.03]

15,000 voxels or features

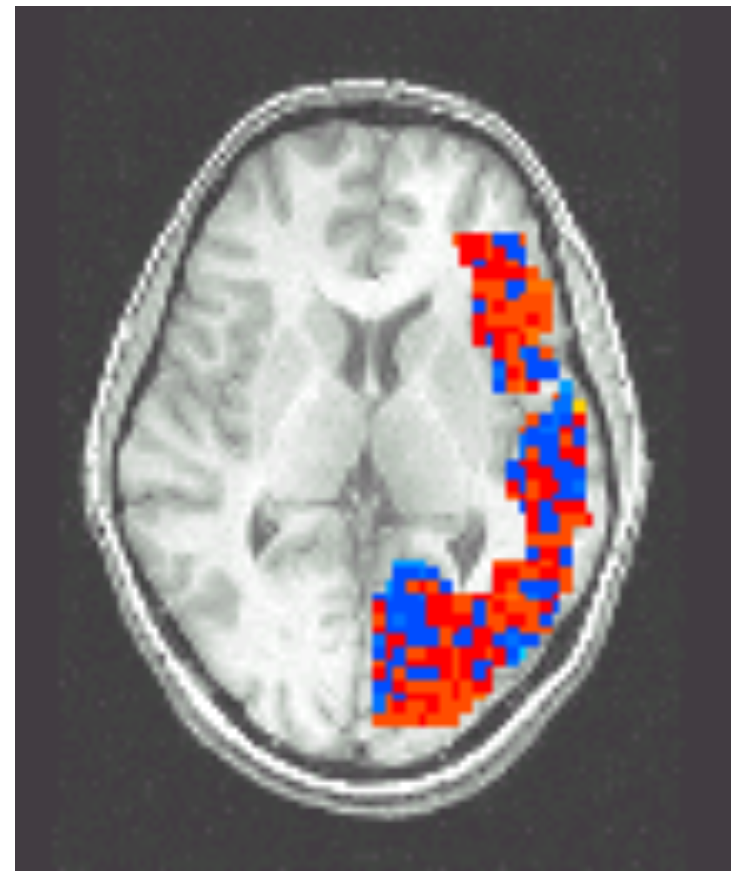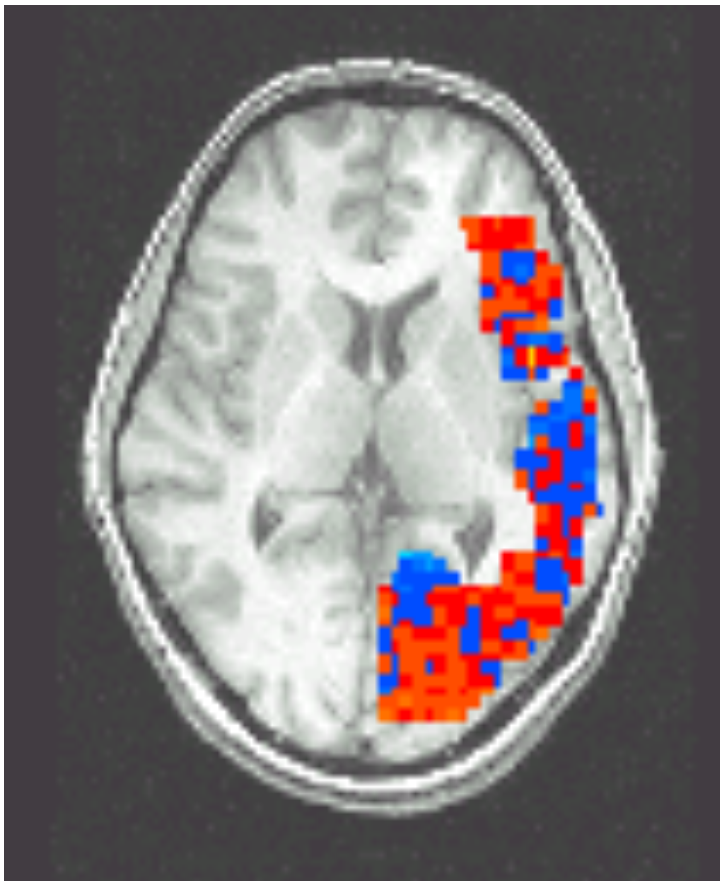10 training examples or subjects per class (12 word categories)

Learned Naïve Bayes Models –
Means for P(BrainActivity | WordCategory)

[Mitchell et al.03]

Pairwise classification accuracy: 85%

People words   −5   0   +5   Animal words

What you should know…

- Optimal decision using Bayes Classifier
- Naïve Bayes classifier
  - What's the assumption
  - Why we use it
  - How do we learn it (MLE, MAP)
  - Why is MAP estimation important