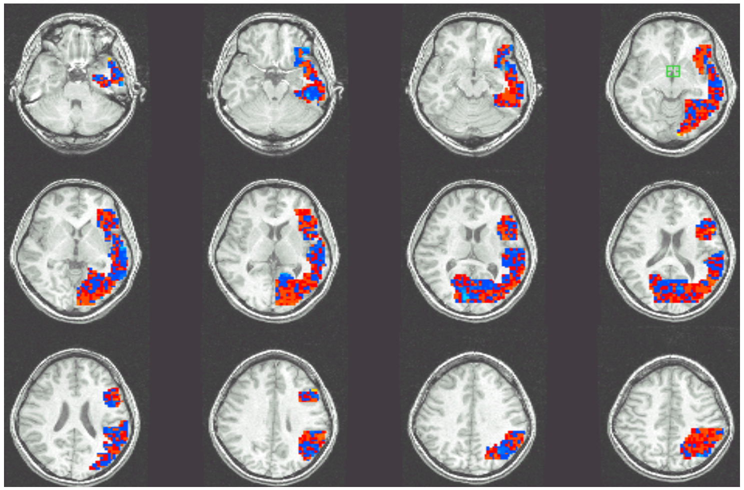
Naïve Bayes
Learning Distributions (MAP)

Aarti Singh

Machine Learning 10-315
Sept 16, 2020

MACHINE LEARNING DEPARTMENT
Carnegie Mellon.
School of Computer Science

Multi-class, multi-dimensional classification – Continuous features



High Stress
Moderate Stress
Low Stress

Input feature vector, X **Label, Y**

We started with a simple case:

label Y is binary (either "Stress" or "No Stress")

X is average brain activity in the "Amygdala"

In general: label Y can belong to K>2 classes

X is multi-dimensional d>1 (average activity in all brain regions)

How many parameters do we need to learn (continuous features)?

Class probability:

$P(Y = y) = p_y$ for all y in H, M, L p_H, p_M, p_L (sum to 1)

K-1 if K labels

Class conditional distribution of features:

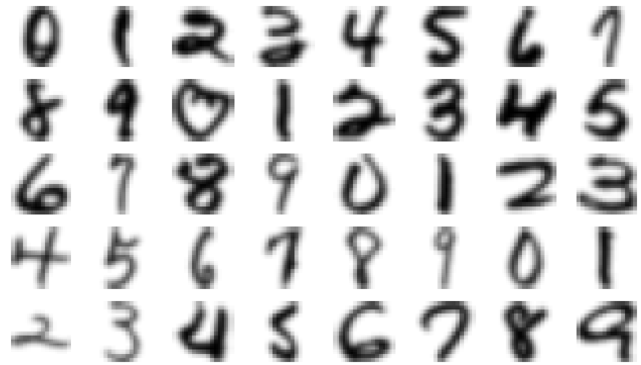
$P(X=x|Y = y) \sim N(\mu_y,\Sigma_y)$ for each y μ_y – d-dim vector

Σ_y - dxd matrix

$Kd + Kd(d+1)/2 = O(Kd^2)$ if d features

Quadratic in dimension d! If d = 256x256 pixels, ~ 13 billion parameters!

Multi-class, multi-dimensional classification - Discrete features



“0”
“1”
…
“9”

Input feature vector, X

Label, Y

Sports
Science
News

Input feature vector, X

Label, Y

How many parameters do we need to learn (discrete features)?

Class probability:

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

K-1 if K labels

Class conditional distribution of (binary) features:

$P(X=x|Y = y)$ ~ For each label y, maintain probability table with 2^d-1 entries

$K(2^d – 1)$ if d binary features

Exponential in dimension d!

What's wrong with too many parameters?

- How many training data needed to learn one parameter (bias of a coin)?

- Need lots of training data to learn the parameters!
  - Training data > number of (independent) parameters

Naïve Bayes Classifier

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

$$X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$$

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

  - More generally:

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \dots \\ X_d \end{bmatrix}$$

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

- If conditional independence assumption holds, NB is optimal classifier! But worse otherwise.

Conditional Independence

- X is **conditionally independent** of Y given Z:

  probability distribution governing X is independent of the value of Y, given the value of Z

$$(\forall x, y, z) P(X = x | Y = y, Z = z) = P(X = x | Z = z)$$

- Equivalent to:

$$P(X, Y \mid Z) = P(X \mid Z) P(Y \mid Z)$$

- e.g., $P(Thunder|Rain, Lightning) = P(Thunder|Lightning)$

  Note: does NOT mean Thunder is independent of Rain

Conditional vs. Marginal Independence

London taxi drivers: A survey has pointed out a positive and significant correlation between the number of accidents and wearing coats. They concluded that coats could hinder movements of drivers and be the cause of accidents. A new law was prepared to prohibit drivers from wearing coats when driving.

Finally another study pointed out that people wear coats when it rains…

Wearing coats is independent of accidents conditioning on the fact that it rained

Naïve Bayes Classifier

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

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

$$\begin{aligned} f_{NB}(\mathbf{x}) &= \arg\max_y \; P(x_1, \ldots, x_d \mid y) P(y) \\ &= \arg\max_y \prod_{i=1}^{d} P(x_i | y) P(y) \end{aligned}$$

- How many parameters now?

How many parameters do we need to learn (continuous features)?

Class probability:

$P(Y = y) = p_y$ for all y in H, M, L $\qquad$ p_H, p_M, p_L (sum to 1)

K-1 if K labels

Class conditional distribution of features (using Naïve Bayes assumption):

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

2Kd $\quad$ if d features

Linear instead of Quadratic in dimension d!

How many parameters do we need to learn (discrete features)?

Class probability:

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

K-1 if K labels

Class conditional distribution of (binary) features:

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

Kd if d binary features

Linear instead of Exponential in dimension d!

Naïve Bayes Classifier

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

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

$$\begin{aligned} f_{NB}(\mathbf{x}) &= \arg\max_y \; P(x_1, \ldots, x_d \mid y) P(y) \\ &= \arg\max_y \prod_{i=1}^{d} P(x_i | y) P(y) \end{aligned}$$

- Has fewer parameters, and hence requires fewer training data, even though assumption may be violated in practice

Naïve Bayes Algo – Discrete features

- Training Data $\{(X^{(j)}, Y^{(j)})\}_{j=1}^{n}$ $\quad X^{(j)} = (X_1^{(j)}, \ldots, X_d^{(j)})$
- Maximum Likelihood Estimates
  - For Class probability
    $$\widehat{P}(y) = \frac{\{\#j : Y^{(j)} = y\}}{n}$$
  - For class conditional distribution
    $$\frac{\widehat{P}(x_i, y)}{\widehat{P}(y)} = \frac{\{\#j : X_i^{(j)} = x_i, Y^{(j)} = y\}/n}{\{\#j : Y^{(j)} = y\}/n}$$
- NB Prediction for test data $\quad X = (x_1, \ldots, x_d)$

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

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'}
  - $\widehat{P}$(X_1= a | Y = b) = 0
- Thus, no matter what the values X_2,…,X_d take:

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

- 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" data

Assume priors

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

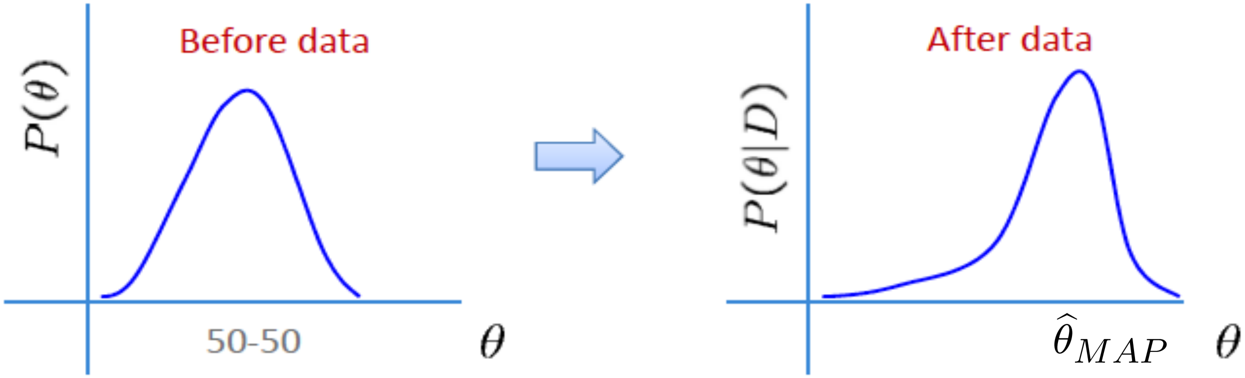
$$\widehat{P}(X_i = a|Y = b) = \frac{\{\#j : X_i^{(j)} = a, Y^{(j)} = b\} + mQ(X_i = a, Y = b)}{\{\#j : Y^{(j)} = b\} + \underbrace{mQ(Y = b)}_{\text{\# virtual examples with Y = b}}}$$

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 (before seeing data D) distribution P(θ) for parameters θ



- Choose value that maximizes a posterior distribution P(θ|D) of parameters θ

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

How to choose prior distribution?

- $P(\theta)$
  - Prior knowledge about domain e.g. unbiased coin $P(\theta) = 1/2$
  - A mathematically convenient form e.g. "conjugate" prior

    If $P(\theta)$ is conjugate prior for $P(D|\theta)$, then Posterior has same form as prior

$$\text{Posterior} = \text{Likelihood} \times \text{Prior}$$

$$P(\theta|D) = P(D|\theta) \times P(\theta)$$

e.g.	Beta	Bernoulli	Beta	θ = bias
	Gaussian	Gaussian	Gaussian	θ = mean μ (known Σ)
	inv-Wishart	Gaussian	inv-Wishart	θ = cov matrix Σ (known μ)

MAP estimation for Bernoulli r.v.

Choose θ that maximizes a posterior probability

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

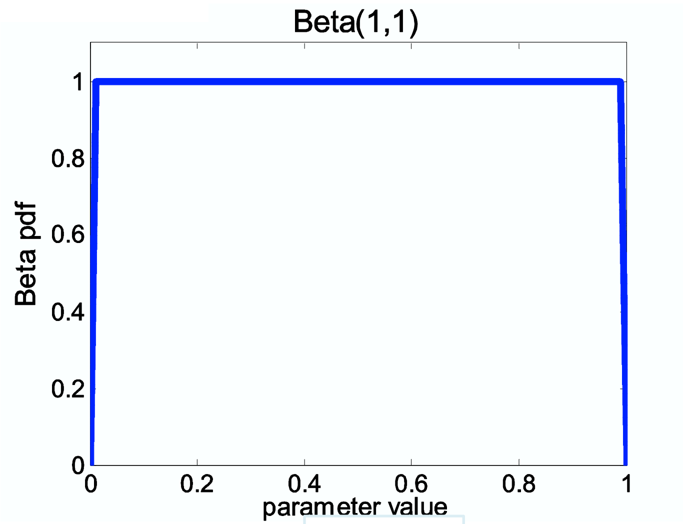
MAP estimate of probability of head (using Beta conjugate prior):

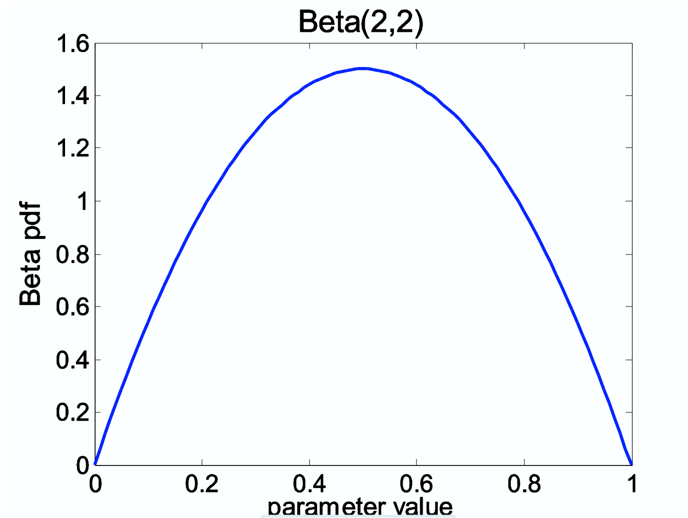
$$P(\theta) \sim Beta(\beta_H, \beta_T)$$

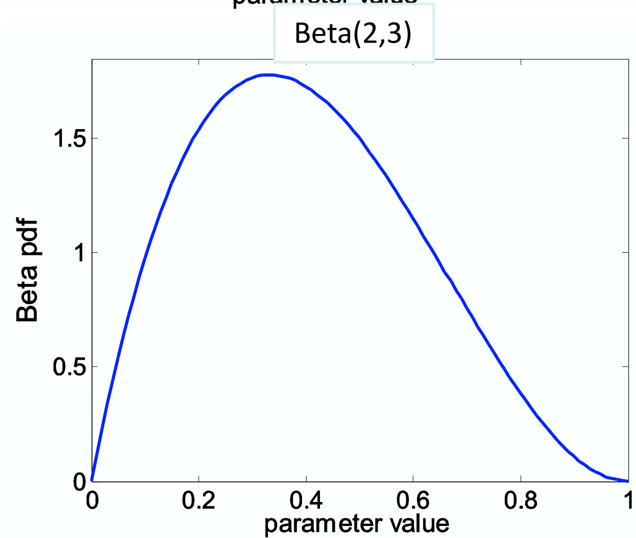
Beta distribution

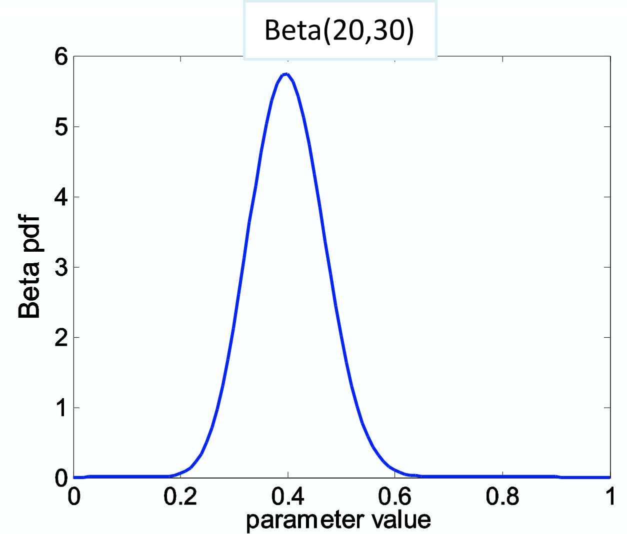
$Beta(\beta_H, \beta_T)$

More concentrated as values of β_H, β_T increase









MAP estimation for Bernoulli r.v.

Choose θ that maximizes a posterior probability

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

MAP estimate of probability of head (using Beta conjugate prior):

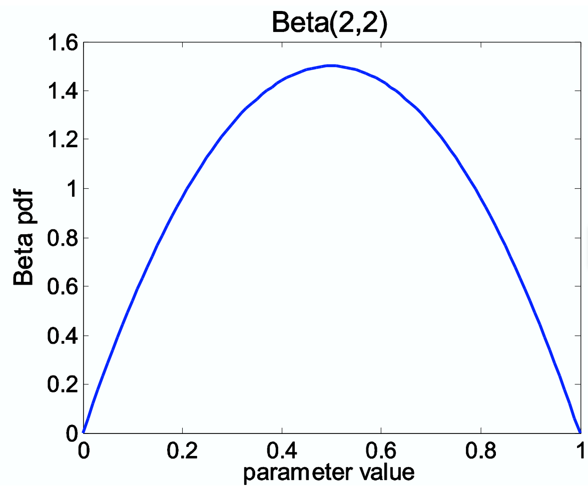
$$P(\theta) \sim Beta(\beta_H, \beta_T)$$

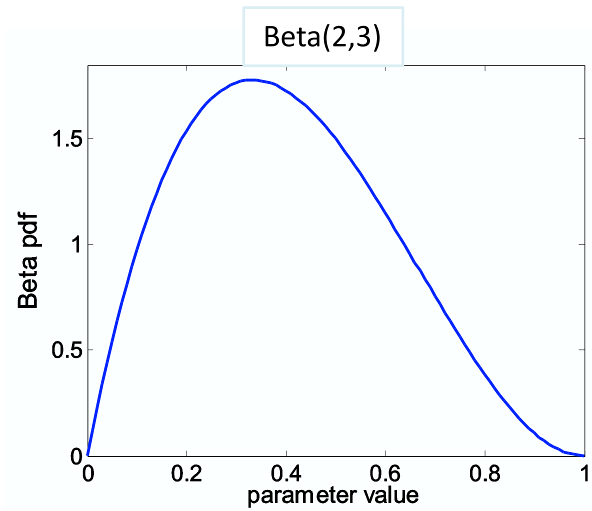
$$P(\theta|D) \sim Beta(\beta_H + \alpha_H, \beta_T + \alpha_T)$$

Count of H/T simply get added to parameters

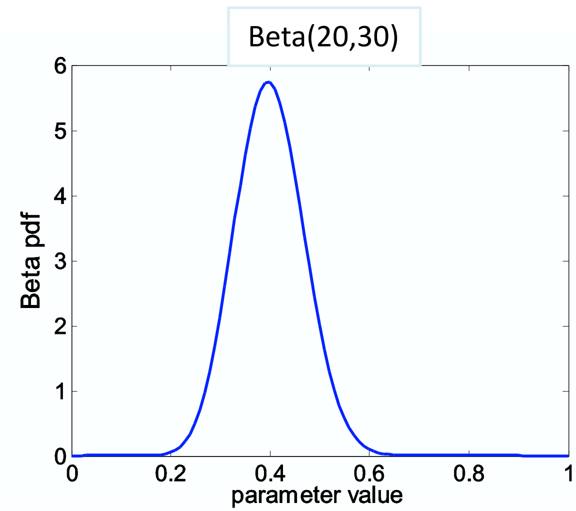
Beta conjugate prior

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





After observing 1 Tail



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

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

MAP estimate of probability of head:

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

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

Count of H/T simply get added to parameters

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

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

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

When is MAP same as MLE?

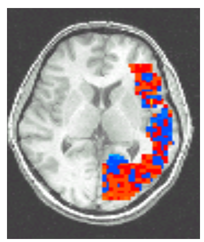
Back to Naïve Bayes (continuous features)

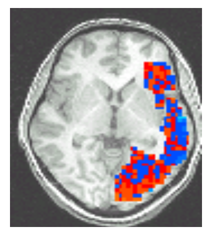
Naïve Bayes with continuous features

Training Data:

Each input represented as a vector of **brain activity values at the d pixels (features)**

Input, X





... n scans

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \dots \\ X_d \end{bmatrix}$$

Label, Y — High stress — Low stress — ... n labels

Gaussian Naïve Bayes model:

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

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

Naïve Bayes Algo – continuous features

- Training Data $\{(X^{(j)}, Y^{(j)})\}_{j=1}^{n}$ $\quad X^{(j)} = (X_1^{(j)}, \ldots, X_d^{(j)})$
- Maximum Likelihood Estimates
  - For Class probability

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

  - For class conditional distribution

$$\hat{P}(x_i|y) = N(\underbrace{\hat{\mu_i}^{(y)}, \hat{\sigma_i}^{2(y)}}_{\text{MLE estimates}})$$

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

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

Naïve Bayes Algo – continuous features

Maximum likelihood estimates:

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

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

y class

i^{th} pixel in j^{th} training image

j^{th} 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: Wishart Distribution

Learned Gaussian Naïve Bayes Model

Means for P(BrainActivity | WordCategory)

[Mitchell et al.03]

Pairwise classification accuracy: 85%

People words

Animal words