Classification - Naïve Bayes with MLE, MAP

Aarti Singh

Co-instructor: Barnabas Poczos

Machine Learning 10-401
Jan 21, 2016

Naïve Bayes Classifier

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

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

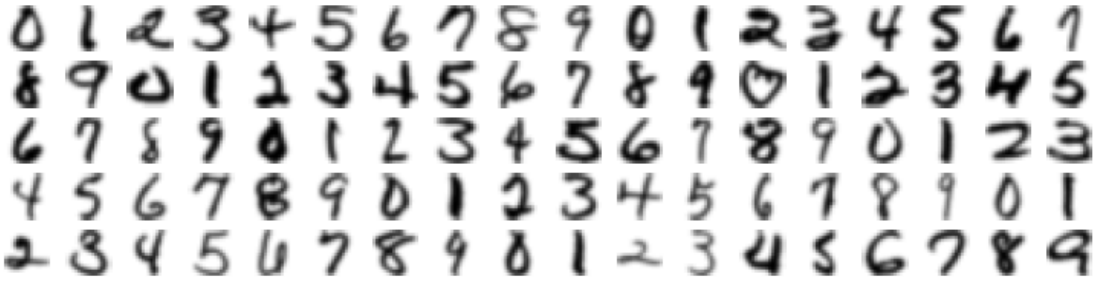
$$\begin{aligned} f_{NB}(\mathbf{x}) &= \arg\max_y \; P(x_1, \ldots, x_d \mid y) P(y) \\ &= \arg\max_y \prod_{i=1}^{d} \; P(x_i | y) P(y) \end{aligned}$$

- Has fewer parameters, and hence requires fewer training data

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 $p_0, p_1, \ldots, p_9$ (sum to 1)

Akin to roll of a dice (Multinomial)

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

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

MLE for Bernoulli, Multinomial

Bernoulli(θ)

P(Heads) = θ, P(Tails) = 1-θ

$$\hat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T}$$

"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)

$$\hat{\theta}_{MLE} = \hat{p}_{1,MLE}, \ldots, \hat{p}_{6,MLE}$$

$$\hat{p}_{y,MLE} = \frac{\alpha_y}{\sum_y \alpha_y}$$

"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 $\hat{P}(y) = \frac{\{\#j : Y^{(j)} = y\}}{n}$
  - For class conditional distribution

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

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

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

Issues with Naïve Bayes

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

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

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

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

- 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

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

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

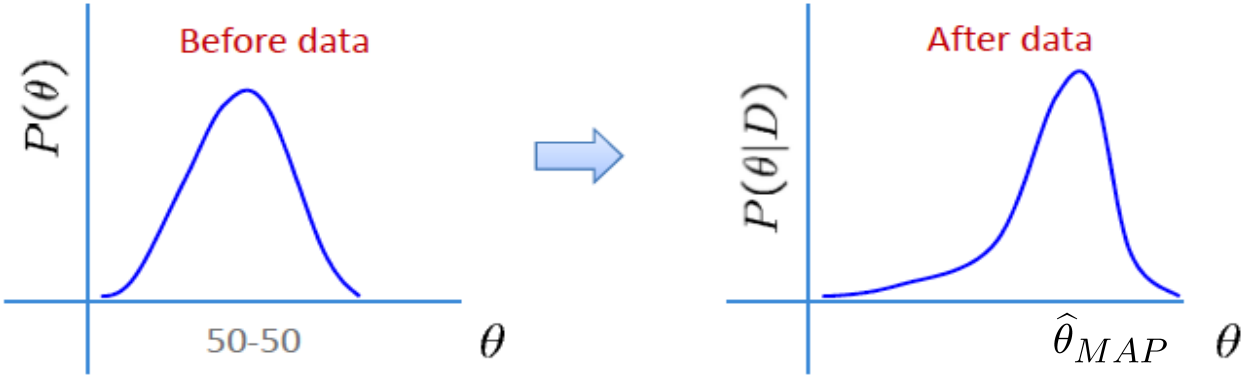
virtual examples with Y = b

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

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

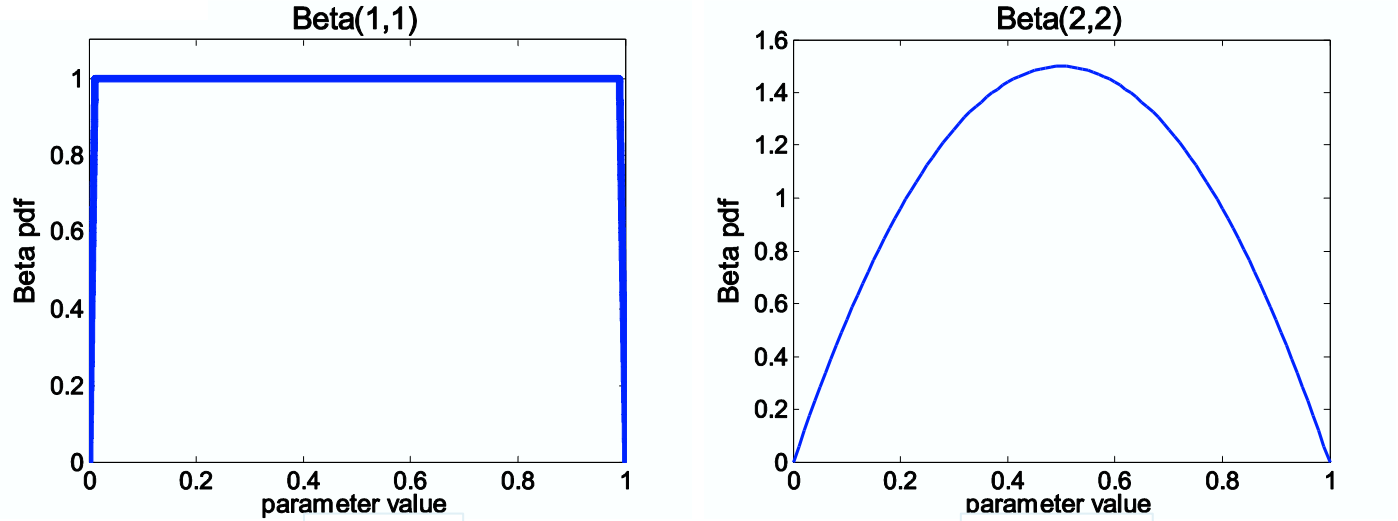
MAP estimate of probability of head:

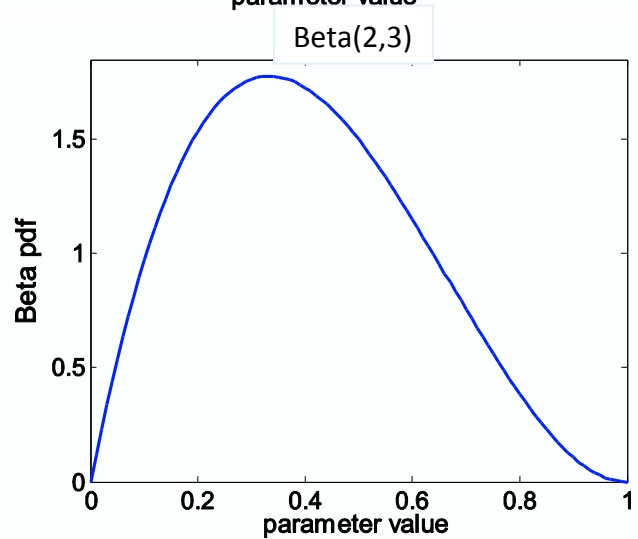
$$P(\theta) \sim Beta(\beta_H, \beta_T)$$

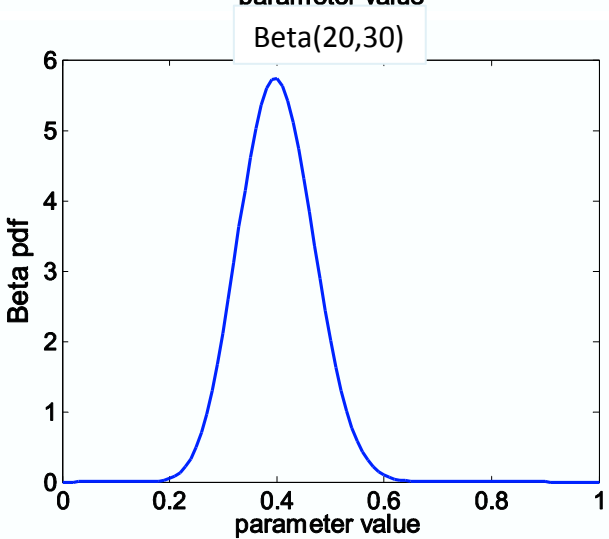
Beta distribution

$Beta(\beta_H, \beta_T)$

More concentrated as values of β_H, β_T increase







MAP estimation for Bernoulli r.v.

Choose θ that maximizes a posterior probability

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

MAP estimate of probability of head:

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

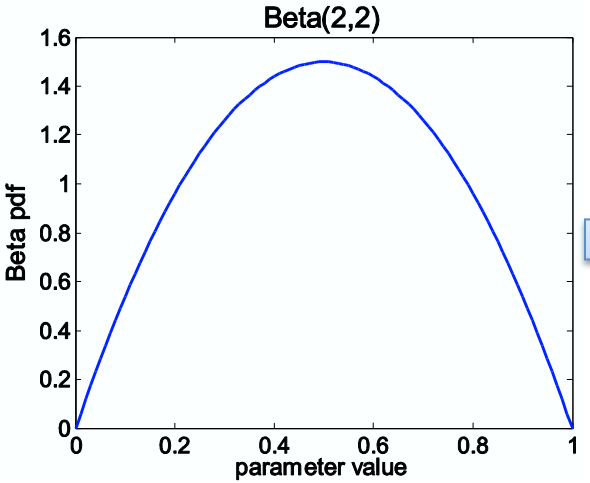
$$P(\theta|D) \sim Beta(\beta_H + \alpha_H, \beta_T + \alpha_T)$$

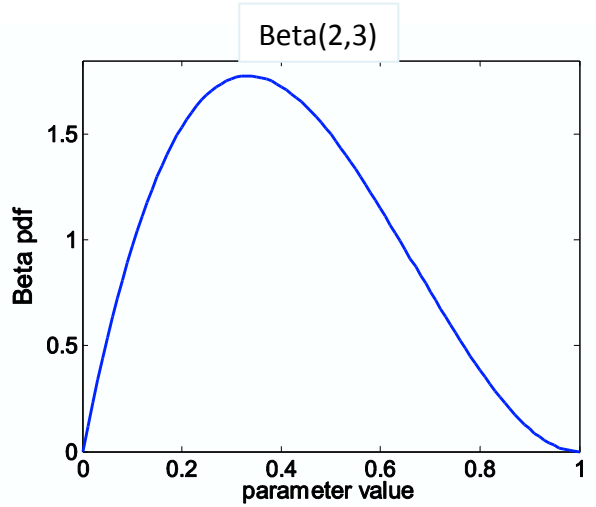
Count of H/T simply get added to parameters

Beta conjugate prior

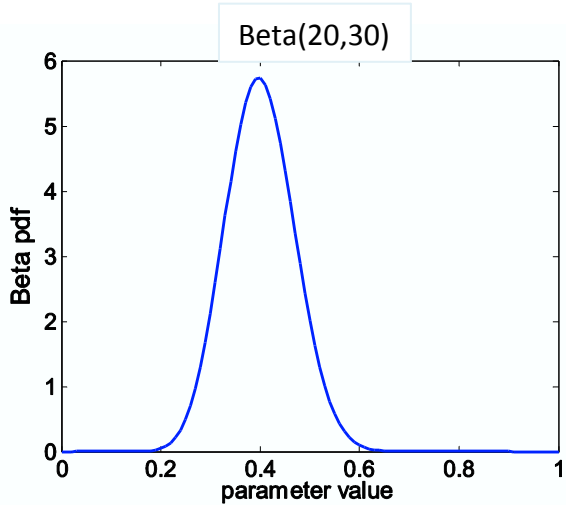
$$P(\theta) \sim Beta(\beta_H, \beta_T)$$

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





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

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

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?

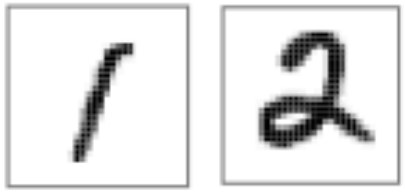
How to learn parameters from data? MLE, MAP

(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  … n greyscale images

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

Label, Y 1 2 … n labels

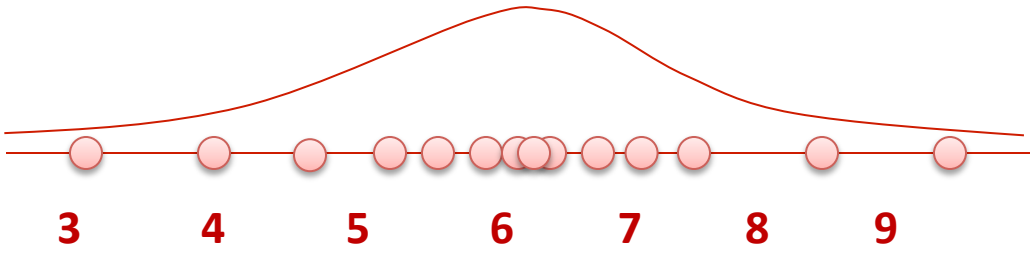
Gaussian Naïve Bayes model:

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

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

Gaussian distribution

Data, D =



- 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

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

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

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

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

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

Note: MLE for the variance of a Gaussian is **biased**

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

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

Naïve Bayes Algo – continuous 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 $\hat{P}(y) = \frac{\{\#j : Y^{(j)} = y\}}{n}$
  - 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 $X = (x_1, \ldots, x_d)$

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

Naïve Bayes Algo – continuous features

Maximum likelihood estimates:

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

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

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

y class

j^{th} training image

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

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

MAP estimation for Gaussian r.v.

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

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

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

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

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

- Variance σ^2: Wishart Distribution

Applications of Naïve Bayes

- Text classification

  Documents ➔ Topic

  Emails ➔ Spam

  HW1

- Image classification

  Hand-written characters ➔ Digit, Letter

  Brain scans ➔ words, objects person is reading/ watching