MLE vs. MAP

Aarti Singh

Machine Learning 10-701/15-781
Sept 15, 2010

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?

MAP using Conjugate Prior

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

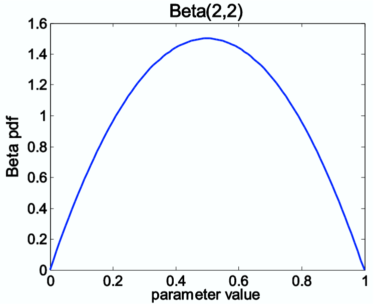
Coin flip problem

Likelihood is ~ Binomial

$$P(\mathcal{D} \mid \theta) = \theta^{\alpha_H}(1-\theta)^{\alpha_T}$$

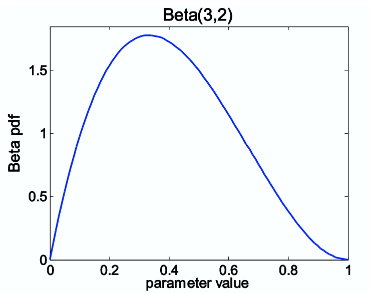
If prior is Beta distribution,

$$P(\theta) \propto \theta^{\beta_H - 1}(1-\theta)^{\beta_T - 1} \sim Beta(\beta_H, \beta_T)$$



Then posterior is Beta distribution

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



For Binomial, conjugate prior is Beta distribution.

MLE vs. MAP

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

What if we toss the coin too few times?

- You say: Probability next toss is a head = 0
- Billionaire says: You're fired! …with prob 1 ☺

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

- Beta prior equivalent to extra coin flips **(regularization)**
- As $n \longrightarrow \infty$, prior is "forgotten"
- **But, for small sample size, prior is important!**

Bayesians vs. Frequentists

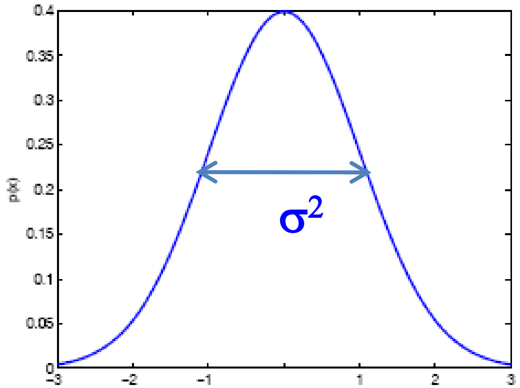
You are no good when sample is small

You give a different answer for different priors

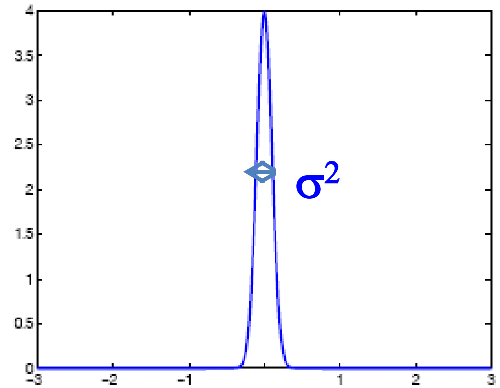
What about continuous variables?

- Billionaire says: If I am measuring a continuous variable, what can you do for me?
- **You say: Let me tell you about Gaussians…**

$$P(x \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} = N(\mu, \sigma^2)$$

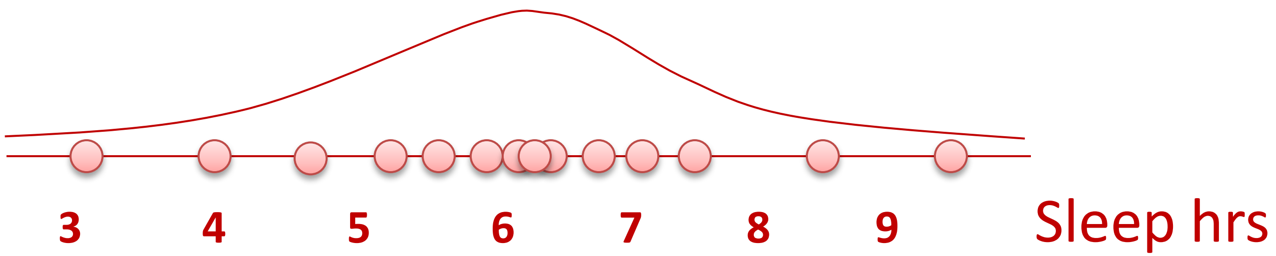


$\mu=0$



$\mu=0$

Gaussian distribution

Data, D = 

- Parameters: μ – mean, σ^2 - variance
- Sleep hrs are **i.i.d.**:
  - **Independent** events
  - **Identically distributed** according to Gaussian distribution

Properties of Gaussians

- affine transformation (multiplying by scalar and adding a constant)
  - $X \sim \mathcal{N}(\mu,\sigma^2)$
  - $Y = aX + b \rightarrow Y \sim \mathcal{N}(a\mu+b,a^2\sigma^2)$

- Sum of Gaussians
  - $X \sim \mathcal{N}(\mu_X,\sigma^2_X)$
  - $Y \sim \mathcal{N}(\mu_Y,\sigma^2_Y)$
  - $Z = X+Y \rightarrow Z \sim \mathcal{N}(\mu_X+\mu_Y, \sigma^2_X+\sigma^2_Y)$

MLE for Gaussian mean and variance

MLE for Gaussian mean and variance

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

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

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

– Expected result of estimation is **not** true parameter!

– Unbiased variance estimator:

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

MAP for Gaussian mean and variance

- Conjugate priors
  - Mean: Gaussian prior
  - Variance: Wishart Distribution

- Prior for mean:

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

MAP for Gaussian Mean

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

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

(Assuming known variance σ^2)

MAP under Gauss-Wishart prior - Homework

What you should know…

- Learning parametric distributions: form known, parameters unknown
  - Bernoulli (θ, probability of flip)
  - Gaussian (μ, mean and σ^2, variance)
- MLE
- MAP

What loss function are we minimizing?

- Learning distributions/densities – Unsupervised learning
- **Task:** Learn $P(X;\theta) \equiv \text{Learn } \theta$ (know form of P, except θ)
- **Experience:** D = $\{X_i\}_{i=1}^n \sim P(X;\theta)$
- **Performance:**

$$\max_\theta P(D|\theta)$$

$$= \min_\theta - \log P(D|\theta)$$

$$= \min_\theta \frac{1}{n} \sum_{i=1}^{n} \underbrace{- \log P(X_i|\theta)}_{\text{loss}(X_i,\theta)}$$

Negative log Likelihood loss

Recitation Tomorrow!

- Linear Algebra and Matlab
- Strongly recommended!!
- Place: NSH 1507 (Note: change from last time)
- Time: 5-6 pm

Leman

Bayes Optimal Classifier

Aarti Singh

Machine Learning 10-701/15-781
Sept 15, 2010

Classification

Goal: Construct a **predictor** $f : X \to Y$ to minimize a risk (performance measure) $R(f)$

Sports
Science
News

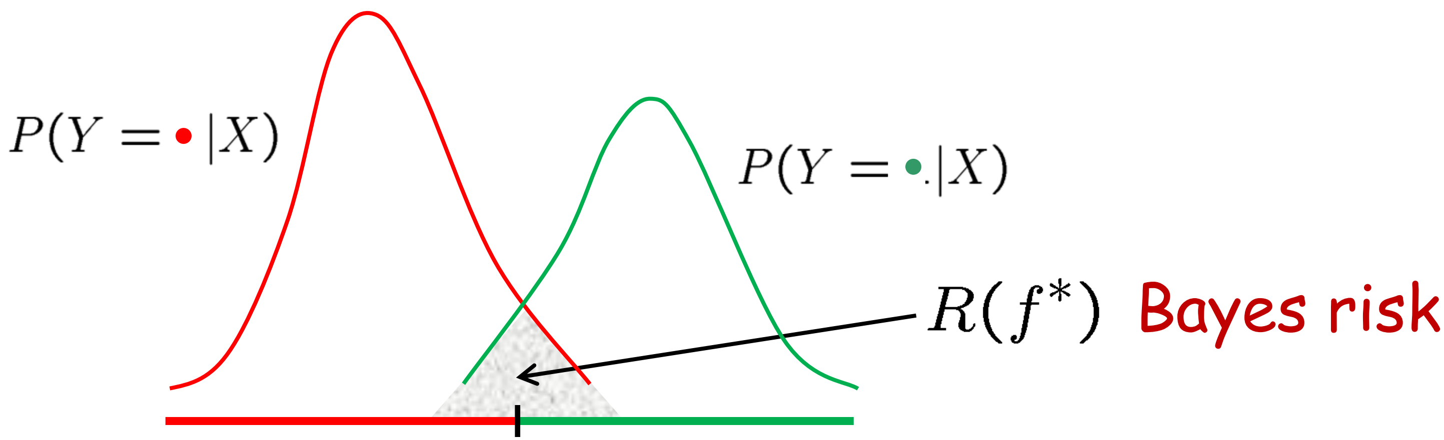
Features, X **Labels, Y**

$$R(f) = P(f(X) \neq Y)$$ **Probability of Error**

Optimal Classification

Optimal predictor:
(Bayes classifier)

$$f^* = \arg\min_f P(f(X) \neq Y)$$



$$f^*(x) = \arg\max_{Y=y} P(Y = y|X = x)$$

- Even the optimal classifier makes mistakes R(f*) > 0
- Optimal classifier depends on **unknown** distribution P_{XY}

Optimal Classifier

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

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

Optimal classifier:

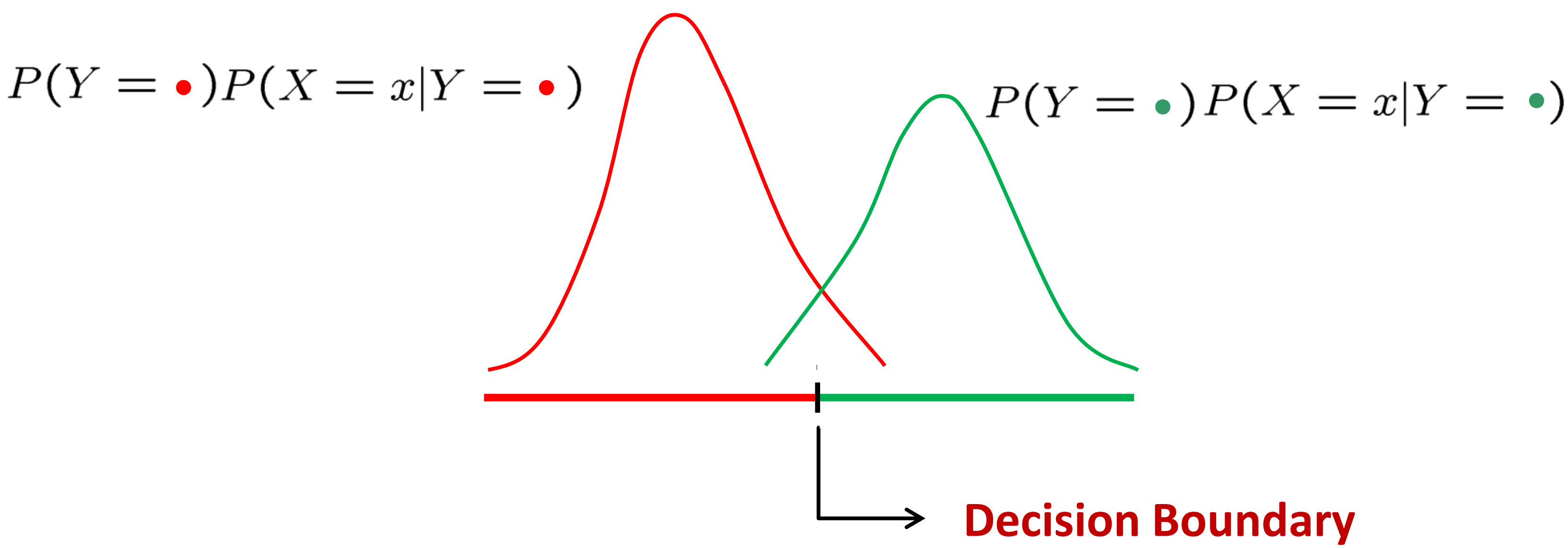
$$f^*(x) = \arg\max_{Y=y} P(Y = y|X = x)$$

$$= \; \arg\max_{Y=y} \underbrace{P(X = x|Y = y)}_{\text{Class conditional density}} \underbrace{P(Y = y)}_{\text{Class prior}}$$

Example Decision Boundaries

- Gaussian class conditional densities (1-dimension/feature)

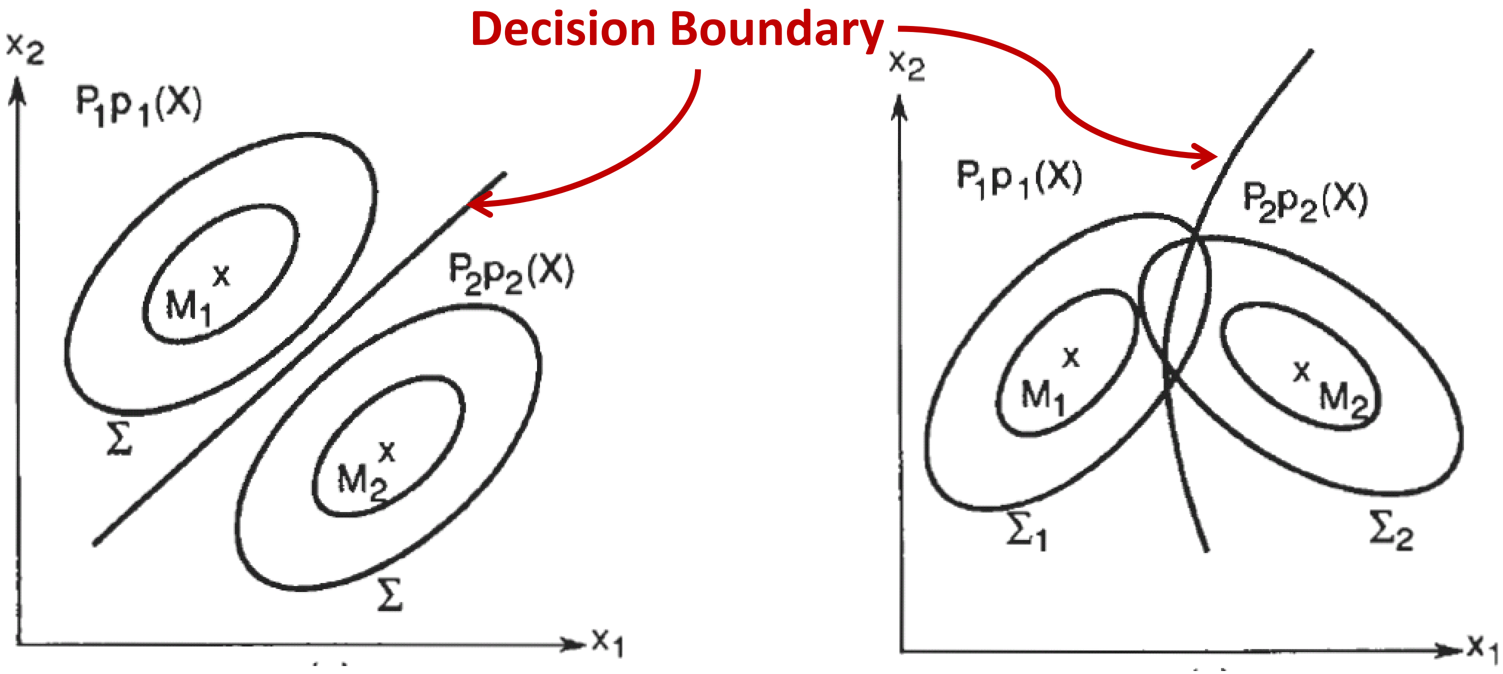
$$P(X = x|Y = y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x-\mu_y)^2}{2\sigma_y^2}\right)$$



Example Decision Boundaries

- Gaussian class conditional densities (2-dimensions/features)

$$P(X = x|Y = y) = \frac{1}{\sqrt{2\pi|\Sigma_y|}} \exp\left(-\frac{(x-\mu_y)\Sigma_y^{-1}(x-\mu_y)'}{2}\right)$$



Learning the Optimal Classifier

Optimal classifier:

$$f^*(x) = \arg\max_{Y=y} P(Y = y|X = x)$$

$$= \arg\max_{Y=y} \underbrace{P(X = x|Y = y)}_{\text{Class conditional density}} \underbrace{P(Y = y)}_{\text{Class prior}}$$

Need to know Prior P(Y = y) for all y

Likelihood P(X=x|Y = y) for all x,y

Learning the Optimal Classifier

Task: Predict whether or not a picnic spot is enjoyable

Training Data: $X = (X_1 \quad X_2 \quad X_3 \quad \ldots \quad \ldots \quad X_d) \qquad Y$

n rows ↕

Sky	Temp	Humid	Wind	Water	Forecst	EnjoySpt
Sunny	Warm	Normal	Strong	Warm	Same	Yes
Sunny	Warm	High	Strong	Warm	Same	Yes
Rainy	Cold	High	Strong	Warm	Change	No
Sunny	Warm	High	Strong	Cool	Change	Yes

Lets learn P(Y|X) – how many parameters?

Prior: P(Y = y) for all y **K-1 if K labels**

Likelihood: P(X=x|Y = y) for all x,y **$(2^d - 1)K$ if d binary features**

Learning the Optimal Classifier

Task: Predict whether or not a picnic spot is enjoyable

Training Data: $X = (X_1 \quad X_2 \quad X_3 \quad \ldots \quad \ldots \quad X_d) \quad Y$

n rows

Sky	Temp	Humid	Wind	Water	Forecst	EnjoySpt
Sunny	Warm	Normal	Strong	Warm	Same	Yes
Sunny	Warm	High	Strong	Warm	Same	Yes
Rainy	Cold	High	Strong	Warm	Change	No
Sunny	Warm	High	Strong	Cool	Change	Yes

Lets learn P(Y|X) – how many parameters?

$2^d K - 1$ (K classes, d binary features)

Need $n \gg 2^d K - 1$ number of training data to learn all parameters