Learning theory: Maximum Likelihood, Bayesian Learning, Model Selection

Shikui Tu
Shanghai Jiao Tong University
2021-03-30

Outline

- Recall from the previous lectures
- Maximum Likelihood (ML) learning
- Bayesian learning, Maximum A Posterior (MAP)
- Model Selection

From minimizing sum of square distances to finding maximum likelihood

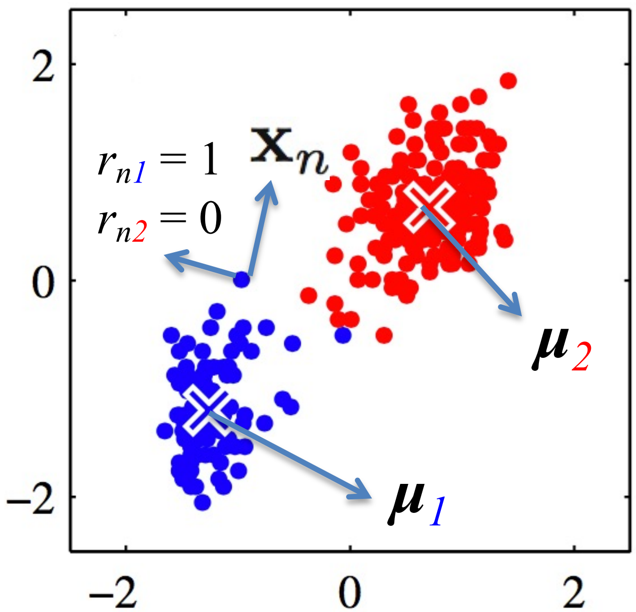
$$X = \{x_1, ..., x_N\}$$
$$\pi = \{\pi_1, ..., \pi_K\}$$
$$\mu = \{\mu_1, \ldots, \mu_K\}$$
$$\Sigma = \{\Sigma_1, ..., \Sigma_K\}$$

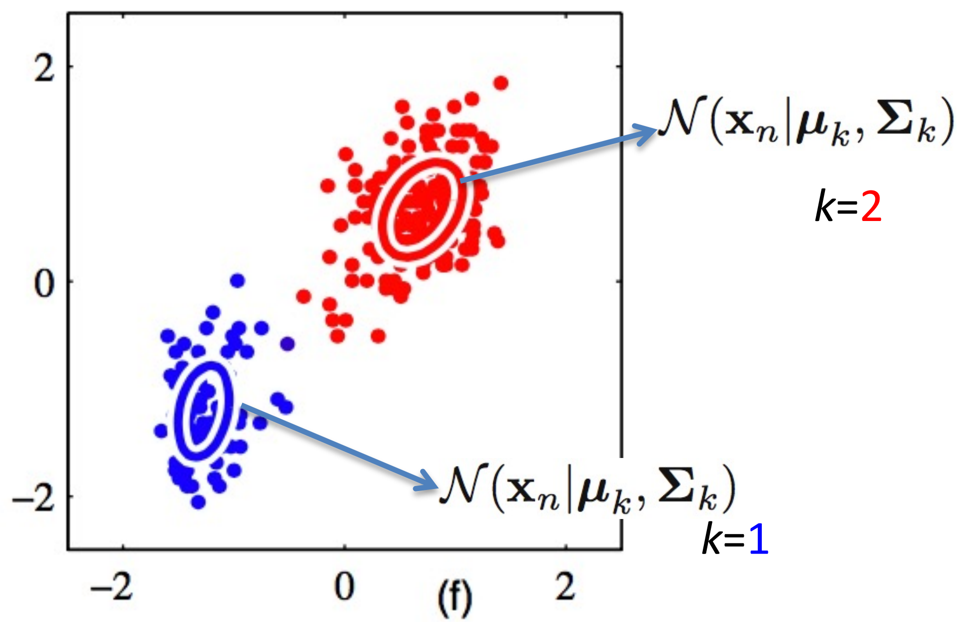
minimize

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2$$

maximize likelihood

$$p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi})$$





Remember: **The closer the distance, the more likely the probability.**

The General EM Algorithm

Given a joint distribution $p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$ over observed variables $\mathbf{X}$ and latent variables $\mathbf{Z}$, governed by parameters $\boldsymbol{\theta}$, the goal is to maximize the likelihood function $p(\mathbf{X}|\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$.

1. Choose an initial setting for the parameters $\boldsymbol{\theta}^{\text{old}}$.
2. **E step** Evaluate $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}})$.
3. **M step** Evaluate $\boldsymbol{\theta}^{\text{new}}$ given by

$$\boldsymbol{\theta}^{\text{new}} = \arg\max_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}})$$

   where

$$\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}) \ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}).$$

4. Check for convergence of either the log likelihood or the parameter values. If the convergence criterion is not satisfied, then let

$$\boldsymbol{\theta}^{\text{old}} \leftarrow \boldsymbol{\theta}^{\text{new}}$$

   and return to step 2.

Summary for the EM algorithm for GMM

- Does it find the global optimum?
  - No, like K-means, EM only finds the nearest local optimum and the optimum depends on the initialization

- GMM is more general then K-means by considering mixing weights, covariance matrices, and soft assignments.

- Like K-means, it does not tell you the best K.

Outline

- Recall from the previous lectures
- **Maximum Likelihood (ML) learning**
- Bayesian learning, Maximum A Posterior (MAP)
- Model Selection

An example

- If flipping a coin a few times, and get

- What is the probability it will fall with the head up?

You may say: 3/5

Because ….

Bernoulli distribution

The dataset $D = \{ x_t \}, t=1,\ldots,N,\ x_t \in \{H, T\}$

$D =$ 

$$P(x = Head) = \theta$$

$$P(x = Tail) = 1 - \theta$$

Flipping coins are **i.i.d.**, i.e., independent identically distributed according to Bernoulli distribution

Question: What is the parameter θ that maximizes the probability of observed data?

Maximum Likelihood Estimation

- Choose parameter θ that maximizes the probability of observed data

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

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

$$= \arg\max_{\theta} \prod_{i:X_i=H} \theta \prod_{i:X_i=T} (1-\theta) \qquad \text{Identically distributed}$$

$$= \arg\max_{\theta} \; \underbrace{\theta^{\alpha_H}(1-\theta)^{\alpha_T}}_{J(\theta)}$$

$$\widehat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T}$$ = 3/5 "Frequency of heads"

Number of heads   Number of tails

Outline

- Recall from the previous lectures
- Maximum Likelihood (ML) learning
- **Bayesian learning, Maximum A Posterior (MAP)**
- Model Selection

Bayesian Learning

- Bayes rule

$$P(\Theta|X) = \frac{P(X|\Theta)P(\Theta)}{P(X)}$$

$P(X|\Theta)$: likelihood of data X given parameter Θ

$P(\Theta)$: prior distribution over the parameter Θ

$P(X)$: marginal distribution of data X

Bayes, Thomas (1763) An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, **53:370-418**

- Prior distribution
  - Represents expert knowledge
  - Uninformative priors: Uniform distribution
  - Conjugate priors: Closed-form representation of posterior, P(θ) and P(θ|D) have the same form

Bayesian learning

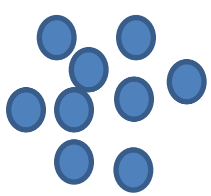
- Maximum A Posteriori (MAP)

$$\max_{\Theta} p(\Theta|X)$$

Equivalent to:

$$\log p(X,\Theta) = \log p(X|\Theta) + \log p(\Theta)$$

Consider a simple example:



$$p(x|\Theta) = G(x|\mu,\Sigma)$$

$$p(\mu) = G(\mu|\mu_0,\sigma_0^2)$$

When is MAP the same as MLE?

- 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

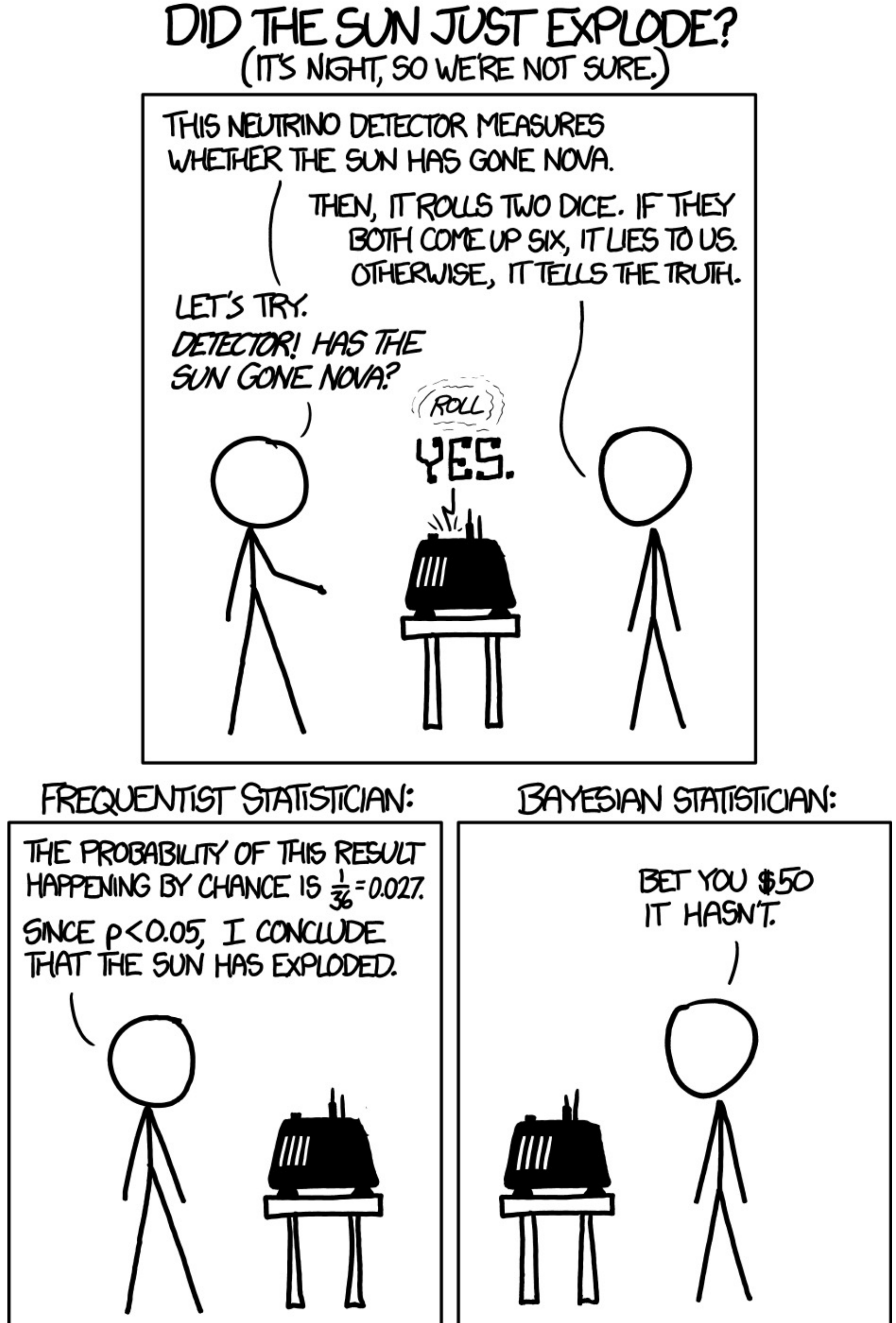
$$\begin{aligned}\widehat{\theta}_{MAP} &= \arg\max_{\theta} P(\theta|D) \\ &= \arg\max_{\theta} P(D|\theta)P(\theta)\end{aligned}$$

Bayesians vs Frequentists

You are no good when sample is small

You give a different answer for different priors

From Machine Learning 10-701/15-781 @CMU-2012


DID THE SUN JUST EXPLODE?
(IT'S NIGHT, SO WE'RE NOT SURE.)
THIS NEUTRINO DETECTOR MEASURES WHETHER THE SUN HAS GONE NOVA.
THEN, IT ROLLS TWO DICE. IF THEY BOTH COME UP SIX, IT LIES TO US. OTHERWISE, IT TELLS THE TRUTH.
LET'S TRY.
DETECTOR! HAS THE SUN GONE NOVA?
ROLL
YES.
FREQUENTIST STATISTICIAN:
THE PROBABILITY OF THIS RESULT HAPPENING BY CHANCE IS $\frac{1}{36} = 0.027$.
SINCE $p < 0.05$, I CONCLUDE THAT THE SUN HAS EXPLODED.
BAYESIAN STATISTICIAN:
BET YOU $50 IT HASN'T.

Thank you!