MLE and MAP

GU 4241

Statistical Machine Learning

Xiaofei Shi

Theoretical foundation: probability

- In order to translate our task into formal mathematical problem, we need the language of

  Probability : the study of uncertainty

COLUMBIA UNIVERSITY

A brief introduction to probability

- Random variables
  refer to an event whose status is unknown:
  - A = "the stock price of google is going to increase by 0.1% tomorrow": binary
  - A = "the app you use for food delivery" : discrete
  - A = "the chance of snow in NYC tomorrow" : continuous
- The set of all possible outcomes
  - All of the possible outcomes a random variable can take

Columbia University

Probability

A variety of useful facts can be derived from just three axioms:

1. $0 \leq P(A) \leq 1$
2. $P(true) = 1$, $P(false) = 0$
3. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Columbia University

Joint probability

P(A, B)

If we assume independence, then P(A, B) = P(A) P(B)

Snow tomorrow	Snow today
1	1
0	0
1	0
1	1
0	1
1	1
0	0
0	0

Columbia University

Joint probability

P(A, B)

If we assume independence, then P(A, B) = P(A) P(B)

P[snow tomorrow] = ½
P[snow today] = ½
P[snow today and tomorrow] = ?

Snow tomorrow	Snow today
1	1
0	0
1	0
1	1
0	1
1	1
0	0
0	0

Columbia University

Joint probability

P(A, B)

If we assume independence, then P(A, B) = P(A) P(B)

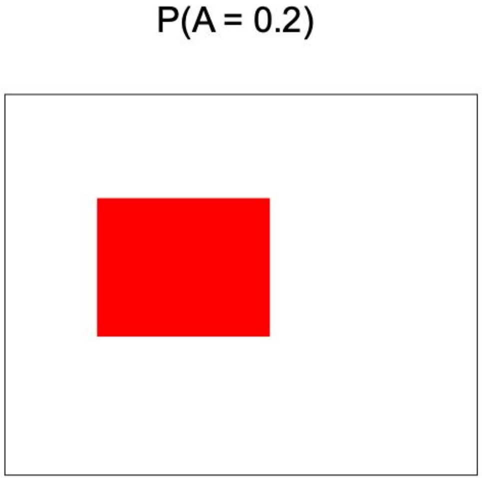
P[snow tomorrow] = ½
P[snow today] = ½
P[snow today and tomorrow] = ⅜

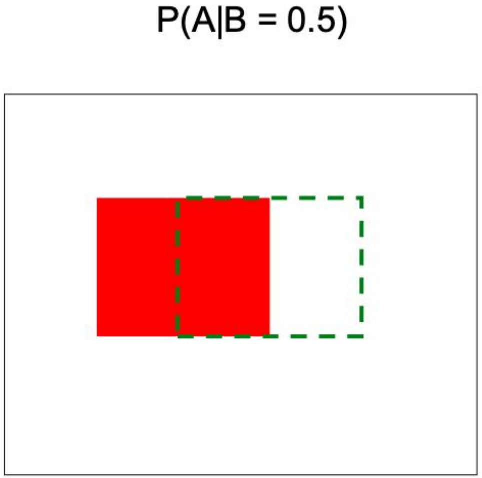
Snow tomorrow	Snow today
1	1
0	0
1	0
1	1
0	1
1	1
0	0
0	0

Columbia University

Conditional probability

P(A | B): The fraction of cases where A is true if B is true





Conditional probability

- In some cases, given knowledge of one or more random variables we can improve upon our prior belief of another random variable
- P[snow tomorrow] = ½
  P[snow tomorrow | snow today] = ¾
  P[no snow tomorrow | snow today] = ¼

Snow tomorrow	Snow today
1	1
0	0
1	0
1	1
0	1
1	1
0	0
0	0

COLUMBIA UNIVERSITY

Chain rule

- The joint probability can be calculated in terms of conditional probability:

$$P(A,B) = P(A|B)\,P(B)$$

- Together with Bayes rule (which is actually derived from it) this is one of the most powerful rules in probabilistic reasoning

Columbia University

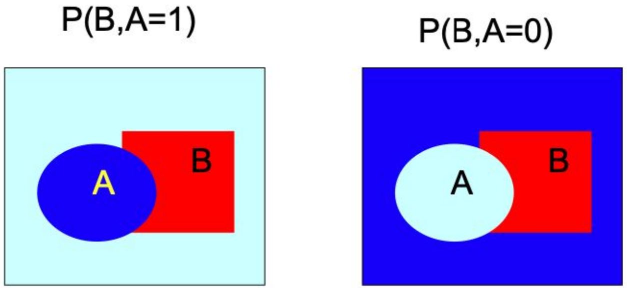
Bayes rule

- Derive from chain rule
- One of the most important rules for this class

Often it would be useful to derive the rule a bit further:

$$P(A|B)=\frac{P(B|A)P(A)}{P(B)}=\frac{P(B|A)P(A)}{\sum_A P(B|A)P(A)}$$

This results from:
$P(B) = \sum_A P(B,A)$



An example

- Suppose you have a coin, if I flip it, what’s the probability it will fall with the head up?
- You might want to flip the coin several times

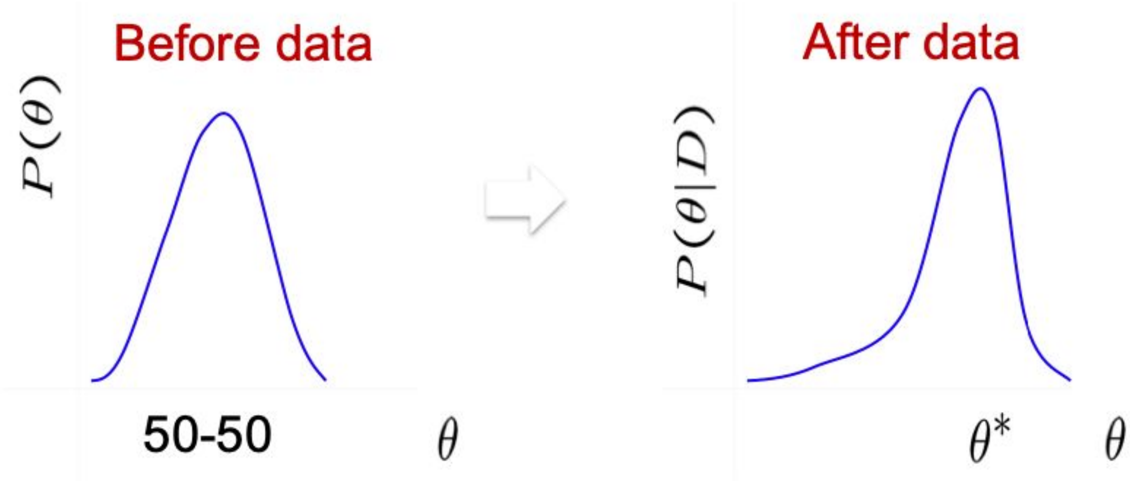
An example

- Suppose you have a coin, if I flip it, what’s the probability it will fall with the head up?
- You might want to flip the coin several times

- The probability is ⅗ because frequency of heads in all flips
- Would you bet money on this estimation?

What about your prior knowledge?

- Rather than estimating a single parameter, we obtain a distribution over possible values of this parameter



COLUMBIA UNIVERSITY

Bayesian learning

- Use Bayes rule:

$$P(\theta \mid \mathcal{D}) = \frac{P(\mathcal{D} \mid \theta) P(\theta)}{P(\mathcal{D})}$$

- Or equivalently:

$$\underbrace{P(\theta \mid \mathcal{D})}_{\text{posterior}} \propto \underbrace{P(\mathcal{D} \mid \theta)}_{\text{likelihood}} \underbrace{P(\theta)}_{\text{prior}}$$

Columbia University

Prior distribution

- Beliefs in an event in the absence of any other information
- Source of prior:

  - Represents expert knowledge (philosophical approach)

  - Simple posterior form (engineer's approach)
- Uninformative priors:

  - Uniform distribution

  - inappropriate distribution
- Conjugate priors:
  - closed-form representation of posteriors

Columbia University

Conjugate prior

Eg. 1 Coin flip problem

Likelihood given Bernoulli model:

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

Conjugate prior

Eg. 1 Coin flip problem

Likelihood given Bernoulli model:

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

If prior is Beta distribution,

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

Then posterior is Beta distribution

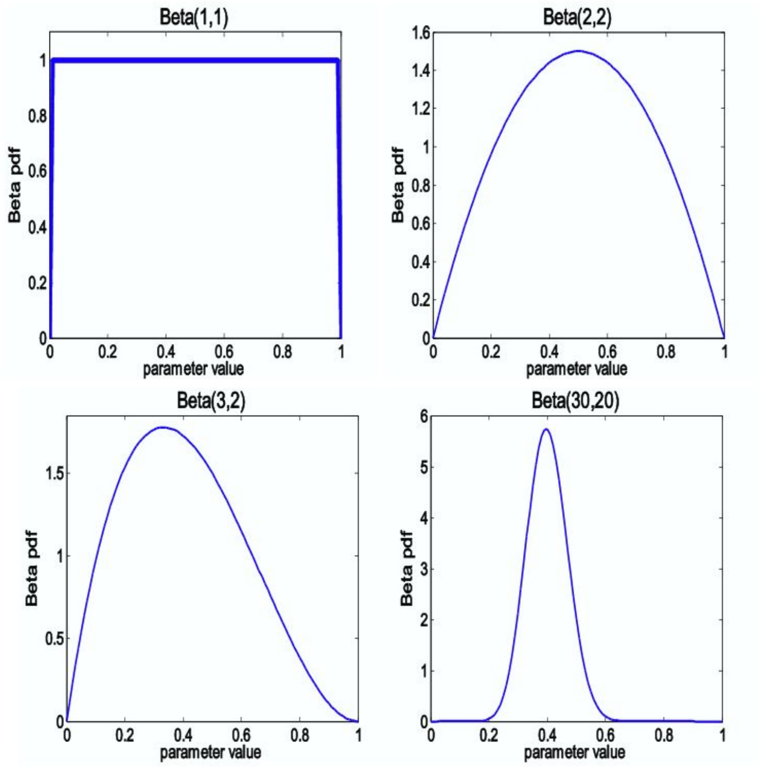
$$P(\theta|D) \sim Beta(\beta_H + \alpha_H, \beta_T + \alpha_T)$$

COLUMBIA UNIVERSITY

Conjugate prior

Beta prior:

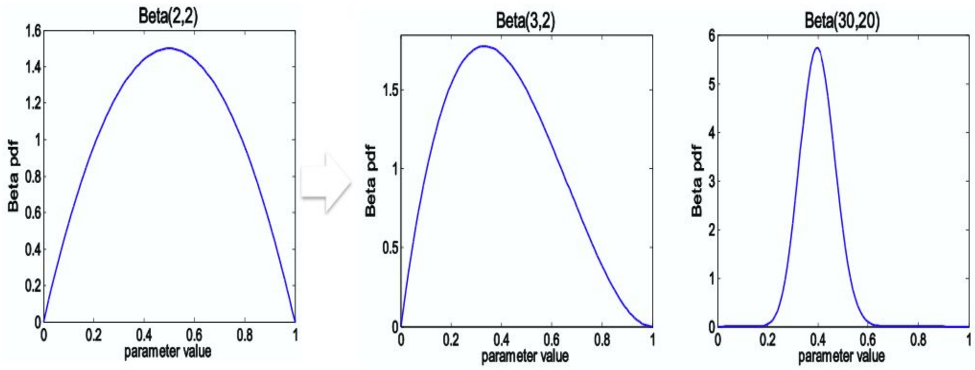
$Beta(\beta_H, \beta_T)$ More concentrated as values of β_H, β_T increase



Conjugate prior

Posterior:

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



As $n = \alpha_H + \alpha_T$ increases

As we get more samples, effect of prior is “washed out”

COLUMBIA UNIVERSITY

Conjugate prior:

- Gaussian prior + Gaussian sample distribution → Gaussian posterior
- Beta prior + Bernoulli sample distribution → Beta posterior
- Gamma prior + exponential sample distribution → Gamma posterior
- Dirichlet prior + multinomial sample distribution → Dirichlet posterior

Posterior distribution

- The approach seen so far is what is known as a Bayesian approach
- Prior information encoded as a distribution over possible values of parameter
- Using the Bayes rule, we can get an updated posterior distribution over parameters

Maximum likelihood principle (MLE)

Data likelihood: $P(D \mid M) = q^{n_1}(1-q)^{n_2}$

We would like to find: $\arg\max_q q^{n_1}(1-q)^{n_2}$

Or more generally, $\hat{P}(\text{dataset} \mid M) = \hat{P}(x_1 \wedge x_2 \ldots \wedge x_n \mid M) = \prod_{k=1}^{n} \hat{P}(x_k \mid M)$

- Our goal is to determine the values for the parameters in M
- We can do this by maximizing the probability of generating the observed samples

An example: Coin flips

MLE

Likelihood: $\theta^{\alpha_H}(1-\theta)^{\alpha_T} =: \mathcal{L}(\theta) \qquad \alpha_H + \alpha_T = N$

$\hat{\theta}_{MLE} =$

MLE

- $\frac{\partial \mathcal{L}(\theta)}{\partial \theta} \Rightarrow \alpha_H \theta^{\alpha_H - 1}(1-\theta)^{\alpha_T} - \alpha_T \theta^{\alpha_H}(1-\theta)^{\alpha_T - 1} = 0$

$\ell(\theta) = -\ln \mathcal{L}(\theta) = -\alpha_H \ln\theta - \alpha_T \ln(1-\theta)$

- $\frac{\partial \ell(\theta)}{\partial \theta} = -\frac{\alpha_H}{\theta} + \frac{\alpha_T}{1-\theta} = 0 \Rightarrow \hat{\theta}_{MLE} = \frac{\alpha_H}{N}$

minimize $\underbrace{-\text{log-likelihood}}_{=:\ \text{loss}}$

COLUMBIA UNIVERSITY

MLE v.s. MAP

- Maximum Likelihood estimation (MLE):

  Choose value that maximizes the probability of observed data

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

- Maximum a posteriori (MAP) estimation:

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

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

Columbia University

MAP: Coin flips

MAP : maximize a posterior

likelihood $\theta^{\alpha_H} (1-\theta)^{\alpha_T}$ prior $\propto \theta^{\beta_H - 1} (1-\theta)^{\beta_T - 1}$

posterior $\theta^{\alpha_H + \beta_H - 1} (1-\theta)^{\alpha_T + \beta_T - 1}$

$\Rightarrow \hat{\theta}_{MAP}$, Loss function $\mathcal{L}(\theta) = -\log$ posterior $= -(\alpha_H + \beta_H - 1)\ln\theta - (\alpha_T + \beta_T - 1)\ln(1-\theta)$

$$\frac{\partial \mathcal{L}(\theta)}{\partial \theta} = -(\alpha_H + \beta_H - 1)\frac{1}{\theta} + (\alpha_T + \beta_T - 1)\frac{1}{1-\theta} = 0$$

$$\Rightarrow (\alpha_T + \beta_T - 1)\,\theta = (\alpha_H + \beta_H - 1)(1-\theta)$$

$$\Rightarrow (\alpha_T + \beta_T + \alpha_H + \beta_H - 2)\,\theta = \alpha_H + \beta_H - 1, \quad \alpha_T + \alpha_H = N$$

$$\Rightarrow \hat{\theta}_{MAP} = \frac{\alpha_H + \beta_H - 1}{N + \beta_T + \beta_H - 2} \rightarrow \frac{\alpha_H}{N} = \hat{\theta}_{MLE}$$

COLUMBIA UNIVERSITY

References

- Kevin Murphy: Machine Learning: A probabilistic perspective, Chapter 2, 5, 6
- Tom Mitchell: Machine Learning, Chapter 6
- Ziv Bar-Joseph, Tom Mitchell, Pradeep Ravikumar and Aarti Singh: CMU 10-701

COLUMBIA UNIVERSITY