Introduction to Machine Learning

Statistics

Mingchen Gao

Computer Science & Engineering
State University of New York at Buffalo
Buffalo, NY, USA
mgao8@buffalo.edu
Slides adapted from Varun Chandola

University at Buffalo
Department of Computer Science and Engineering
School of Engeering and Applied Sciences

Outline

Generative Models

- $\mathbf{X}$, feature vector, represents the data with multiple discrete attributes
- Y represents the class

Most probable class

$$P(Y = c|\mathbf{X} = \mathbf{x}, \boldsymbol{\theta}) \propto P(\mathbf{X} = \mathbf{x}|Y = c, \boldsymbol{\theta})P(Y = c, \boldsymbol{\theta})$$

- $p(\mathbf{x}|y = c, \boldsymbol{\theta})$ - **class conditional density**
- How is the data distributed for each class?

Steps for Learning a Generative Model

- Example: D is a sequence of N binary values (0s and 1s) (coin tosses)
- What is the best distribution that could describe D?
- What is the probability of observing a *head* in future?

Step 1: Choose the form of the model

- Hypothesis Space - All possible distributions
  - Too complicated!!
- Revised hypothesis space - All Bernoulli distributions $(X \sim Ber(\theta), 0 \leq \theta \leq 1)$
  - θ is the hypothesis
  - Still infinite (θ can take infinite possible values)

Compute Likelihood

- Likelihood of D

$$p(D|\theta) = \theta^{N_1}(1-\theta)^{N_0}$$

Maximum Likelihood Estimate

$$\begin{aligned}\hat{\theta}_{MLE} &= \arg\max_{\theta} p(D|\theta) = \arg\max_{\theta} \theta^{N_1}(1-\theta)^{N_0} \\ &= \frac{N_1}{N_0 + N_1}\end{aligned}$$

Compute Likelihood

- Likelihood of D

$$p(D|\theta) = \theta^{N_1}(1-\theta)^{N_0}$$

Maximum Likelihood Estimate

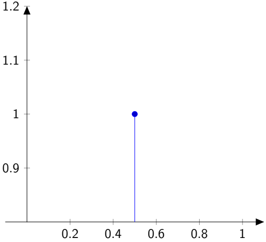
$$\begin{aligned} \hat{\theta}_{MLE} &= \arg\max_{\theta} p(D|\theta) = \arg\max_{\theta} \theta^{N_1}(1-\theta)^{N_0} \\ &= \frac{N_1}{N_0 + N_1} \end{aligned}$$

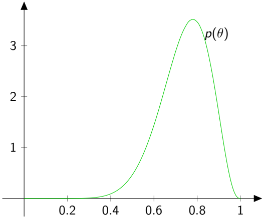
- **We can stop here (MLE approach)**
- Probability of getting a head next:

$$p(x^* = 1|D) = \hat{\theta}_{MLE}$$

Incorporating Prior

- Prior *encodes* our prior belief on θ
- How to set a Bayesian prior?
  1. A point estimate: $\theta_{prior} = 0.5$
  2. A probability distribution over θ (**a random variable**)
     - Which one?
     - For a bernoulli distribution $0 \leq \theta \leq 1$
     - *Beta* Distribution





Beta Distribution as Prior

- Continuous random variables defined between 0 and 1

$$Beta(\theta|a,b) \triangleq p(\theta|a,b) = \frac{1}{B(a,b)}\theta^{a-1}(1-\theta)^{b-1}$$

- a and b are the (hyper-)parameters for the distribution
- $B(a,b)$ is the **beta function**

$$B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

$$\Gamma(x) = \int_0^\infty u^{x-1}e^{-u}du$$

If x is integer

$$\Gamma(x) = (x-1)!$$

- "Control" the shape of the pdf
- **We can stop here as well (prior approach)**

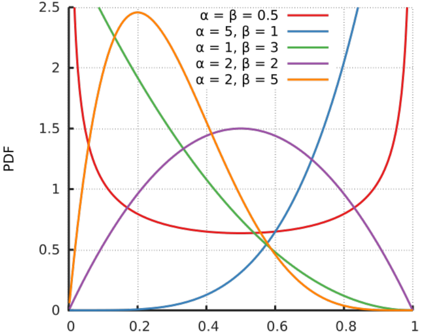
$$p(x^* = 1) = \theta_{prior}$$

Properties of Beta Distribution

$$mean = \frac{a}{a+b}$$

$$mode = \frac{a-1}{a+b-2}$$

$$var = \frac{ab}{(a+b)^2(a+b+1)}$$



Conjugate Priors

- Another reason to choose Beta distribution

$$p(D|\theta) = \theta^{N_1}(1-\theta)^{N_0}$$

$$p(\theta) \propto \theta^{a-1}(1-\theta)^{b-1}$$

- Posterior $\propto$ Likelihood $\times$ Prior

$$\begin{aligned} p(\theta|D) &\propto \theta^{N_1}(1-\theta)^{N_0}\theta^{a-1}(1-\theta)^{b-1} \\ &\propto \theta^{N_1+a-1}(1-\theta)^{N_0+b-1} \end{aligned}$$

- **Posterior has same form as the prior**
- Beta distribution is a conjugate prior for Bernoulli/Binomial distribution

Estimating Posterior

- Posterior

$$\begin{aligned} p(\theta|D) &\propto \theta^{N_1+a-1}(1-\theta)^{N_0+b-1} \\ &= Beta(\theta|N_1+a, N_0+b) \end{aligned}$$

- After observing N trials in which we observe N_1 heads and N_0 tails, we update our belief as:

$$\mathbb{E}[\theta|D] = \frac{a+N_1}{a+b+N}$$

Using Posterior

- We know that posterior over θ is a beta distribution
- MAP estimate

$$
\begin{aligned}
\hat{\theta}_{MAP} &= \arg\max_{\theta} p(\theta | a + N_1, b + N_0) \\
&= \frac{a + N_1 - 1}{a + b + N - 2}
\end{aligned}
$$

- What happens if $a = b = 1$?
- **We can stop here as well (MAP approach)**
- Probability of getting a head next:

$$p(x^* = 1 | D) = \hat{\theta}_{MAP}$$

True Bayesian Approach

- All values of θ are possible
- Prediction on an unknown input (x^*) is given by *Bayesian Averaging*

$$
\begin{aligned}
p(x^* = 1|D) &= \int_0^1 p(x^* = 1|\theta)p(\theta|D)d\theta \\
&= \int_0^1 \theta Beta(\theta|a + N_1, b + N_0) \\
&= \mathbb{E}[\theta|D] \\
&= \frac{a + N_1}{a + b + N}
\end{aligned}
$$

- This is same as using $\mathbb{E}[\theta|D]$ as a point estimate for θ

The Black Swan Paradox

- Why use a *prior*?
- Consider $D = \texttt{tails, tails, tails}$
- $N_1 = 0, N = 3$
- $\hat{\theta}_{MLE} = 0$
- $p(x^* = 1|D) = 0!!$
  - Never observe a heads
  - The *black swan* paradox
- How does the Bayesian approach help?

$$p(x^* = 1|D) = \frac{a}{a + b + 3}$$

Why is MAP Estimate Insufficient?

- MAP is only one part of the posterior
  - θ at which the posterior probability is maximum
  - But is that enough?
  - What about the posterior variance of θ?

$$var[\theta|D] = \frac{(a + N_1)(b + N_0)}{(a + b + N)^2(a + b + N + 1)}$$

  - If variance is high then θ_{MAP} is not trustworthy
  - Bayesian averaging helps in this case

Likelihood

- Why choose one hypothesis over other?
- Avoid **suspicious coincidences**
- Choose concept with higher *likelihood*

$$p(D|h) = \prod_{x \in D} p(x|h)$$

- *Log Likelihood*

$$\log p(D|h) = \sum_{x \in D} \log p(x|h)$$

The Principle of Occam's Razor

- Always choose the simpler explanation
- A general problem-solving philosophy

Finding the Best Hypothesis

Maximum A Priori Estimate

$$\hat{h}_{prior} = \arg\max_h p(h)$$

Maximum Likelihood Estimate (MLE)

$$\begin{aligned}\hat{h}_{MLE} &= \arg\max_h p(D|h) = \arg\max_h \log p(D|h) \\ &= \arg\max_h \sum_{x \in D} \log p(x|h)\end{aligned}$$

Maximum a Posteriori (MAP) Estimate

$$\hat{h}_{MAP} = \arg\max_h p(D|h)p(h) = \arg\max_h (\log p(D|h) + \log p(h))$$

MAP and MLE

- $\hat{h}_{prior}$ - Most likely hypothesis based on prior
- $\hat{h}_{MLE}$ - Most likely hypothesis based on evidence
- $\hat{h}_{MAP}$ - Most likely hypothesis based on posterior

$$\hat{h}_{prior} = \arg\max_h \log p(h)$$

$$\hat{h}_{MLE} = \arg\max_h \log p(D|h)$$

$$\hat{h}_{MAP} = \arg\max_h (\log p(D|h) + \log p(h))$$

Interesting Properties

- As data increases, MAP estimate converges towards MLE
  - Why?
- MAP/MLE are **consistent estimators**
  - If concept is in $\mathcal{H}$, MAP/ML estimates will converge
- If $c \notin \mathcal{H}$, MAP/ML estimates converge to h which is closest possible to the truth

Posterior Predictive Distribution

- New input, x^*
- What is the probability that x^* is also generated by the same concept as D?
  - $P(x^* \in C|x^*, D)$?
- **Option 0:** Treat h^{prior} as the true concept

$$P(x^* \in C|x^*, D) = P(x^* \in h^{prior}|x^*, h^{prior})$$

- **Option 1:** Treat h^{MLE} as the true concept

$$P(x^* \in C|x^*, D) = P(x^* \in h^{MLE}|x^*, h^{MLE})$$

- **Option 2:** Treat h^{MAP} as the true concept

$$P(x^* \in C|x^*, D) = P(x^* \in h^{MAP}|x^*, h^{MAP})$$

- **Option 3:** *Bayesian Averaging*

$$P(x^* \in C|x^*, D) = \sum_h P(x^* \in h|x^*, h)p(h|D)$$

Multivariate Gaussian

- pdf for MVN with d dimensions:

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \triangleq \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]$$

Estimating Parameters of MVN

Problem Statement

Given a set of N **independent and identically distributed** (iid) samples, D, learn the parameters $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ of a Gaussian distribution that generated D.

- MLE approach - maximize log-likelihood
- Result

$$\widehat{\boldsymbol{\mu}}_{MLE} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x_i} \triangleq \bar{\mathbf{x}}$$

$$\widehat{\boldsymbol{\Sigma}}_{MLE} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x_i} - \bar{\mathbf{x}})(\mathbf{x_i} - \bar{\mathbf{x}})^\top$$

References

Chapter 4.1 - 4.2.5, 4.6 Murphy Book