Introduction to Machine Learning

Bayesian Learning

Varun Chandola

Computer Science & Engineering
State University of New York at Buffalo
Buffalo, NY, USA
chandola@buffalo.edu

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

Outline

Generative Models

- Let $\mathbf{X}$ represents the data with multiple discrete attributes
- Y represent the class

Most probable class

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

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

Concept Learning in Number Line

- I give you a set of numbers (training set D) belonging to a concept
- Choose the most likely hypothesis (concept)
- Assume that numbers are between 1 and 100
- Hypothesis Space ($\mathcal{H}$):
  - All powers of 2
  - All powers of 4
  - All even numbers
  - All prime numbers
  - Numbers close to a fixed number (say 12)
  - $\vdots$

Hypothesis Space ($\mathcal{H}$)

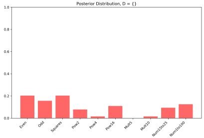
1. Even numbers
2. Odd numbers
3. Squares
4. Powers of 2
5. Powers of 4

Hypothesis Space ($\mathcal{H}$)

6. Powers of 16
7. Multiples of 5
8. Multiples of 10
9. Numbers within 20 ± 5
10. All numbers between 1 and 100

Ready?

- $D = \{\}$



Hypothesis Space ($\mathcal{H}$)

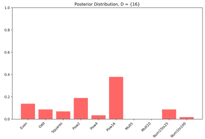
1. Even numbers
2. Odd numbers
3. Squares
4. Powers of 2
5. Powers of 4
6. Powers of 16
7. Multiples of 5
8. Multiples of 10
9. Numbers within 20 ± 5
10. All numbers between 1 and 100

Ready?

Hypothesis Space ($\mathcal{H}$)

1. Even numbers
2. Odd numbers
3. Squares
4. Powers of 2
5. Powers of 4
6. Powers of 16
7. Multiples of 5
8. Multiples of 10
9. Numbers within 20 ± 5
10. All numbers between 1 and 100

- $D = \{16\}$

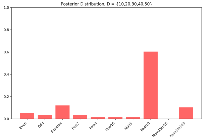


Ready?

Hypothesis Space ($\mathcal{H}$)

1. Even numbers
2. Odd numbers
3. Squares
4. Powers of 2
5. Powers of 4
6. Powers of 16
7. Multiples of 5
8. Multiples of 10
9. Numbers within 20 ± 5
10. All numbers between 1 and 100

- $D = \{20, 30, 40, 50\}$



Ready?

Hypothesis Space ($\mathcal{H}$)

1. Even numbers
2. Odd numbers
3. Squares
4. Powers of 2
5. Powers of 4
6. Powers of 16
7. Multiples of 5
8. Multiples of 10
9. Numbers within 20 ± 5
10. All numbers between 1 and 100

Ready?

Hypothesis Space ($\mathcal{H}$)

1. Even numbers
2. Odd numbers
3. Squares
4. Powers of 2
5. Powers of 4
6. Powers of 16
7. Multiples of 5
8. Multiples of 10
9. Numbers within 20 ± 5
10. All numbers between 1 and 100

- $D = \{1, 4, 16, 64\}$

Computing Likelihood

- Why choose *powers of 4* concept over *even numbers* concept for $D = \{1, 4, 16, 64\}$?
- Avoid **suspicious coincidences**
- Choose concept with higher *likelihood*
- What is the likelihood of above D to be generated using the *powers of 4* concept?
- Likelihood for *even numbers* concept?

Likelihood

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

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

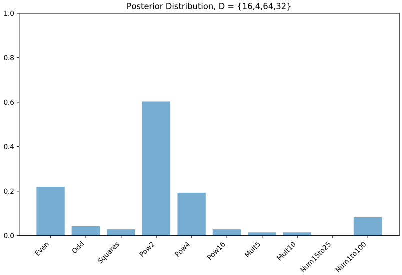
- *Log Likelihood*

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

Bayesian Concept Learning

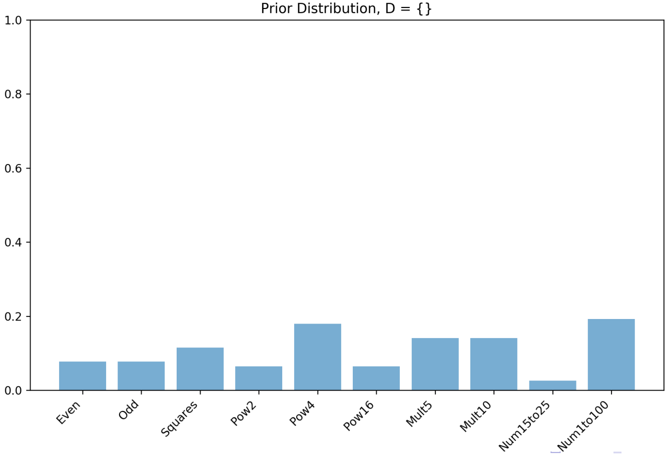
1. Even numbers
2. Odd numbers
3. Squares
4. Powers of 2
5. Powers of 4
6. Powers of 16
7. Multiples of 5
8. Multiples of 10
9. Numbers within 20 ± 5
10. All numbers between 1 and 100

$$D = \{1, 4, 16, 64\}$$



Adding a Prior

- Inside information about the hypotheses
- Some hypotheses are *more likely* apriori
  - May not be the right hypothesis (**prior can be wrong**)



Posterior

- Revised estimates for h after observing evidence (D) and the prior
- *Posterior* $\propto$ *Likelihood* $\times$ *Prior*

$$p(h|D) = \frac{p(D|h)p(h)}{\sum_{h' \in \mathcal{H}} p(D|h')p(h')}$$

	h	Prior	Likelihood	Posterior
1	Even	0.300	1.600e-07	1.403e-04
2	Odd	0.075	0.000e+00	0.000e+00
3	Squares	0.075	1.000e-04	2.192e-02
4	Powers of 2	0.100	4.165e-04	1.217e-01
5	Powers of 4	0.075	3.906e-03	8.562e-01
6	Powers of 16	0.075	0.000e+00	0.000e+00
7	Multiples of 5	0.075	0.000e+00	0.000e+00
8	Multiples of 10	0.075	0.000e+00	0.000e+00
9	Numbers within 20 $\pm$ 5	0.075	0.000e+00	0.000e+00
10	All Numbers	0.075	1.000e-08	2.192e-06

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

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

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

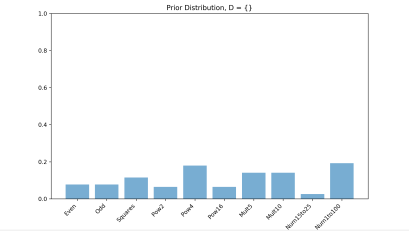
$$\hat{h}_{MAP} = \arg\max_h (\log p(D|h) + \log p(h))$$

Interesting Properties

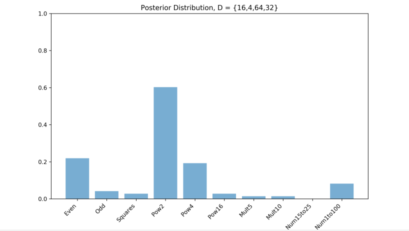
- As data increases, MAP estimate converges towards MLE
  - Why?
- MAP/MLE are **consistent estimators**
  - If concept is in $\mathcal{H}$, MAP/ML estimates will converge
- If $c \notin \mathcal{H}$, MAP/ML estimates converge to h which is closest possible to the truth

From Prior to Posterior via Likelihood

Prior



Posterior



- Objective: To *revise* the prior distribution over the hypotheses after observing data (evidence).

Posterior Predictive Distribution

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

$$P(Y = c|X = x^*, D) = P(X = x^*|c = h^{prior})$$

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

$$P(Y = c|X = x^*, D) = P(X = x^*|c = h^{MLE})$$

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

$$P(Y = c|X = x^*, D) = P(X = x^*|c = h^{MAP})$$

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

$$P(Y = c|X = x^*, D) = \sum_h P(X = x^*|c = h)p(h|D)$$

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

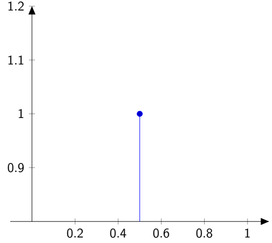
$$\begin{aligned}\hat{\theta}_{MLE} &= \arg\max_{\theta} p(D|\theta) = \arg\max_{\theta} \theta^{N_1}(1-\theta)^{N_0} \\ &= \frac{N_1}{N_0 + N_1}\end{aligned}$$

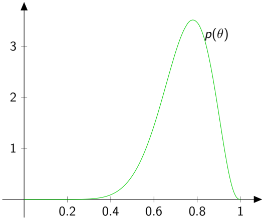
- **We can stop here (MLE approach)**
- Probability of getting a head next:

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

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)**

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

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

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

- We start with a belief that

$$\mathbb{E}[\theta] = \frac{a}{a+b}$$

- After observing N trials in which we observe N_1 heads and N_0 trails, 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

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

- 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

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

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

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

Estimating Posterior

- We need posterior for both $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$

$$p(\boldsymbol{\mu})$$

$$p(\boldsymbol{\Sigma})$$

- What distribution do we need to sample μ?
  - A Gaussian distribution!

$$p(\boldsymbol{\mu}) = \mathcal{N}(\boldsymbol{\mu}|\mathbf{m}_0, \mathbf{V}_0)$$

- What distribution do we need to sample $\boldsymbol{\Sigma}$?
  - An *Inverse-Wishart* distribution.

$$\begin{aligned} p(\boldsymbol{\Sigma}) &= IW(\boldsymbol{\Sigma}|\mathbf{S}, \nu) \\ &= \frac{1}{Z_{IW}}|\boldsymbol{\Sigma}|^{-(\nu+D+1)/2} exp\left(-\frac{1}{2}tr(\mathbf{S}^{-1}\boldsymbol{\Sigma}^{-1})\right) \end{aligned}$$

where,

$$Z_{IW} = |\mathbf{S}|^{-\nu/2}2^{\nu D/2}\Gamma_D(\nu/2)$$

Calculating Posterior

Posterior for $\boldsymbol{\mu}$ - Also a MVN

$$
\begin{aligned}
p(\boldsymbol{\mu}|D, \boldsymbol{\Sigma}) &= \mathcal{N}(\mathbf{m_N}, \mathbf{V_N}) \\
\mathbf{V}_N^{-1} &= \mathbf{V}_0^{-1} + N\boldsymbol{\Sigma}^{-1} \\
\mathbf{m}_N &= \mathbf{V}_N(\boldsymbol{\Sigma}^{-1}(N\bar{\mathbf{x}}) + \mathbf{V}_0^{-1}\mathbf{m}_0)
\end{aligned}
$$

Posterior for $\boldsymbol{\Sigma}$ - Also an Inverse Wishart

$$
\begin{aligned}
p(\boldsymbol{\Sigma}|D, \boldsymbol{\mu}) &= IW(\mathbf{S_N}, \nu_N) \\
\nu_N = \nu_0 + N & \\
\mathbf{S}_N^{-1} &= \mathbf{S}_0 + \mathbf{S}_\mu
\end{aligned}
$$

References