Machine Learning

AIML CLZG565

Bayesian Learning

Raja vadhana P

Assistant Professor,

BITS - CSIS

BITS Pilani

Pilani Campus

Disclaimer and Acknowledgement

- The content for these slides has been obtained from books and various other source on the Internet
- I here by acknowledge all the contributors for their material and inputs.
- I have provided source information wherever necessary
- I have added and modified the content to suit the requirements of the course

Source: Slides of Prof. Chetana, Prof.Seetha, Prof.Sugata, Prof.Vimal, Prof.Monali, Prof. Raja vadhana , Prof.Anita from BITS Pilani , CS109 and CS229 stanford lecture notes, Tom Mitchell, Andrew Ng and many others who made their course materials freely available online

Course Plan

M1	Introduction & Mathematical Preliminaries
M2	Machine Learning Workflow
M3	Linear Models for Regression
M4	Linear Models for Classification
M5	Decision Tree
M6	Instance Based Learning
M7	Support Vector Machine
M8	Bayesian Learning
M9	Ensemble Learning
M10	Unsupervised Learning
M11	Machine Learning Model Evaluation/Comparison

Course Plan

M1	Introduction & Mathematical Preliminaries
M2	Machine Learning Workflow
M3	Linear Models for Regression
M4	Linear Models for Classification
M5	Decision Tree
M6	Instance Based Learning
M7	Support Vector Machine
M8	Bayesian Learning
M9	Ensemble Learning
M10	Unsupervised Learning
M11	Machine Learning Model Evaluation/Comparison

Bayesian Learning
Parameter Estimation

- Where does the **cost** come from? - Logistic regression
- Why least-squares **cost** function, be a reasonable choice? – Linear regression

BITS Pilani

Distribution

Mileage (in kmpl)	*Car Price (in cr)*
9.8	10.48
9.12	1.75
9.5	6.95
10	2.51

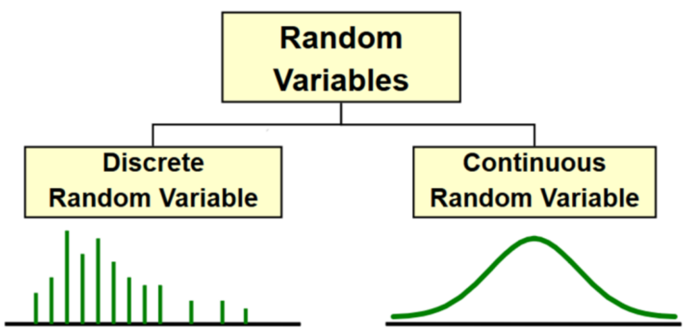
Mileage (in kmpl)	*Car Price (in cr)*
Neutral	High
Less	Low
Neutral	Medium
More	Low

Mileage (in kmpl)	*Car Price (in cr)*
9.8	High
9.12	Low
9.5	High
10	Low

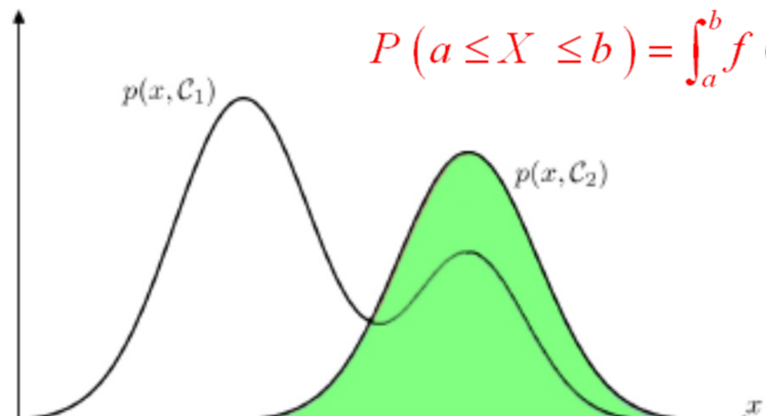
$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N}$$

$$p(Y = y_j | X = x_i) = \frac{n_{ij}}{c_i}$$

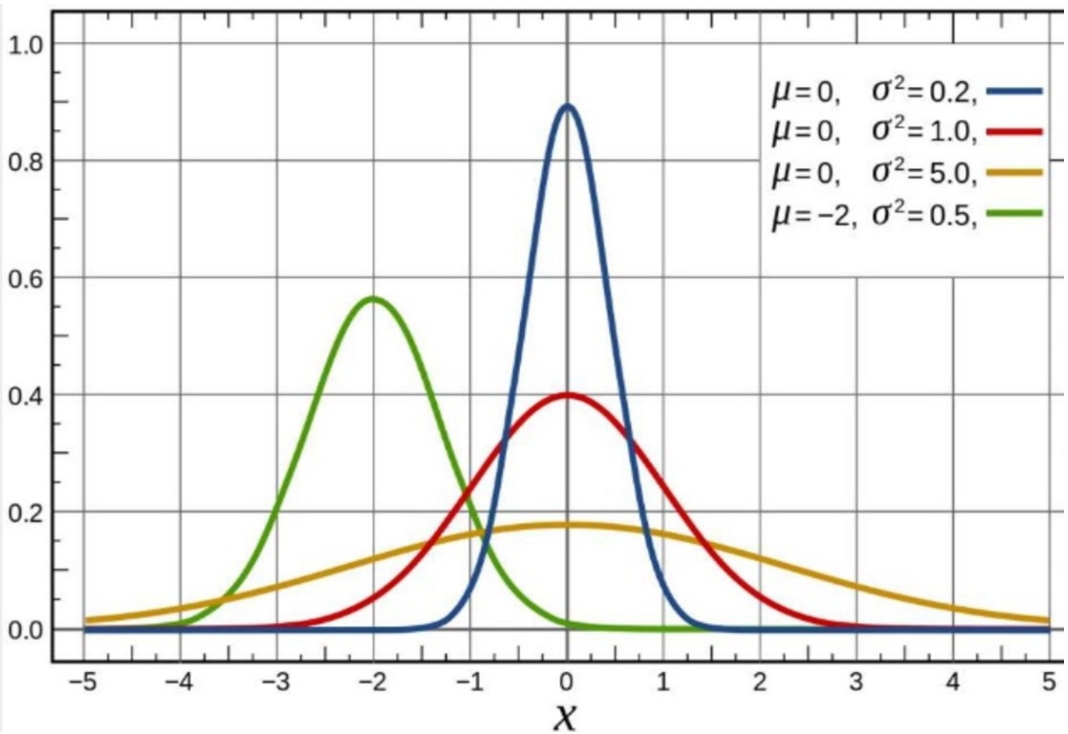
Represents a possible numerical value from a random event



$$P\left(a \le X \le b\right) = \int_a^b f\left(x\right)dx$$



Parameter Estimation



Distribution	Parameters
Bernoulli(p)	$\theta = p$
Poisson(λ)	$\theta = \lambda$
Uniform(a,b)	$\theta = (a, b)$
Normal(μ, σ^2)	$\theta = (\mu, \sigma^2)$
Y = mX + b	$\theta = (m, b)$

θ is the parameter of a distribution.

θ can be a vector of parameters

Distribution = model + parameter θ

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

Find $\boldsymbol{\theta} = (\mathbf{Mean}, \mathbf{SD})$ from the data $\mathbf{X_i}$

ie., (x_i, P(x_i))

Parameter in ML

Mileage (in kmpl)	***Car Price (in cr)***
9.8	10.48
9.12	1.75
9.5	6.95
10	2.51

$$\text{CarPrice} = 8.5 + 0.5\ \text{Mileage} - 1.5\ \text{Mileage}^2$$

$$\text{Parameters} : (\theta_0, \theta_1, \theta_2)$$

Mileage (in kmpl)	***Car Price (in cr)***
9.8	High
9.12	Low
9.5	High
10	Low

$$\text{CarPrice} = \frac{1}{1+e^{-8.5 + 0.5\ \text{Mileage} - 1.5\ \text{Mileage}^2}}$$

Parameter Estimation in ML

Mileage (in kmpl)	*Car Price (in cr)*
9.8	10.48
9.12	1.75
9.5	6.95
10	2.51

CarPrice $= 8.5 + 0.5$ Mileage $- 1.5$ Mileage^2

$y^{(i)} = \theta^T x^{(i)} + \varepsilon^{(i)}$

$\varepsilon^{(i)} \sim N(0, \sigma^2)$

Parameters : $(\theta_0, \theta_1, \theta_2)$

Find $\boldsymbol{\theta} = (\boldsymbol{\theta_0}, \boldsymbol{\theta_1}, \boldsymbol{\theta_2})$ from the data $\mathbf{X_i}$

ie., (Mileage_i, CarPrice_i)

Assumption: Data are__

- IID samples: $X_1 \ldots X_n$ where all X_i are independent and have the same distribution.
- Either same PMF (discrete) or same PDF (continuous)
- $f(X \mid \theta)$ Likelihood of different values of X depends on the values of our parameters θ f( ) is either PDF or PMF

Maximum Likelihood Estimation (MLE)

select that parameters θ that make the observed data the most likely

$f(X_1, X_2 \ldots X_n \mid \boldsymbol{\theta})$

Maximum A Posteriori (MAP)

choose the parameters θ that is the most likely, given the data

$f(\boldsymbol{\theta} \mid X_1, X_2 \ldots X_n)$

Intuition of Bayes Theorem

MAP : f ($\boldsymbol{\theta}$ | X_1, X_2 .....X_n)

MLE: f (X_1, X_2 .....X_n | $\boldsymbol{\theta}$)

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

P(h) = prior probability of hypothesis h

P(D) = prior probability of training data D

P(h|D) = probability of h given D

P(D|h) = probability of D given h

$L(\theta)$, probability of data given parameter θ

After seeing data, posterior belief of θ

Before seeing data, prior belief of θ
e.g. what is distribution over parameters θ

likelihood prior

posterior

$$P(\theta|\text{data}) = \frac{P(\text{data}|\theta)P(\theta)}{P(\text{data})}$$

Maximum Likelihood Estimation (MLE)

1. Determine formula for $LL(\theta)$

2. Differentiate $LL(\theta)$ w.r.t. (each) θ

3. Solve

MLE – Linear Regression Model

Given the noise $\varepsilon^{(i)}$ obeys a Normal distribution each $y^{(i)}$ must also obey a Normal distribution around the true target value

Mileage (in kmpl)	*Car Price (in cr)*
9.8	10.48
9.12	1.75
9.5	6.95
10	2.51

CarPrice $= 8.5 + 0.5$ Mileage $- 1.5$ Mileage^2

$y^{(i)} = \theta^T x^{(i)} + \boldsymbol{\varepsilon^{(i)}}$

$\varepsilon^{(i)} \sim N(0, \sigma^2)$

$$p(\epsilon^{(i)}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(\epsilon^{(i)})^2}{2\sigma^2}\right).$$

Parameters : $(\theta_0, \theta_1, \theta_2)$

Find $\boldsymbol{\theta} = (\boldsymbol{\theta_0}, \boldsymbol{\theta_1}, \boldsymbol{\theta_2})$ from the data $\mathbf{X_i}$

ie., (Mileage$_i$, CarPrice$_i$)

$\mathbf{y^{(i)} \mid x^{(i)}; \theta \sim N(\theta^T x^{(i)}, \sigma^2)}$

$$p(y^{(i)}|x^{(i)};\theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}\right)$$

Maximum Likelihood Estimation (MLE)

select that parameters θ that make the observed data the most likely

f $(X_1, X_2 \; .....X_n \mid \boldsymbol{\theta})$

MLE – Linear Regression Model

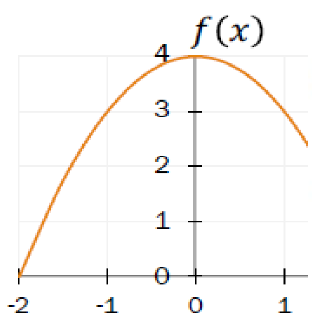
$$p(y^{(i)}|x^{(i)};\theta) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(y^{(i)}-\theta^T x^{(i)})^2}{2\sigma^2}\right)$$

1. Determine formula for $LL(\theta)$
2. Differentiate $LL(\theta)$ w.r.t. (each) θ
3. Solve

$$L(\theta) = \prod_{i=1}^{n} p(y^{(i)} \mid x^{(i)};\theta)$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(y^{(i)}-\theta^T x^{(i)})^2}{2\sigma^2}\right)$$

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, L(\boldsymbol{\theta}) \qquad LL(\theta) = \log L(\theta)$$



$$= \log L(\theta)$$

$$= \log \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(y^{(i)}-\theta^T x^{(i)})^2}{2\sigma^2}\right)$$

$$= \sum_{i=1}^{n} \log \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(y^{(i)}-\theta^T x^{(i)})^2}{2\sigma^2}\right)$$

$$= n\log\frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{\sigma^2}\cdot\frac{1}{2}\sum_{i=1}^{n}(y^{(i)}-\theta^T x^{(i)})^2.$$

Mileage (in kmpl)	*Car Price (in cr)*
9.8	10.48
9.12	1.75
9.5	6.95
10	2.51

MLE answers the questio

observed data have the l

MLE – Linear Regression Model

$$p(y^{(i)}|x^{(i)};\theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}\right)$$

1. Determine formula for $LL(\theta)$

$$\log L(\theta) = n \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{\sigma^2} \cdot \frac{1}{2} \sum_{i=1}^{n} (y^{(i)} - \theta^T x^{(i)})^2$$

2. Differentiate $LL(\theta)$ w.r.t. (each) θ

$$= -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y^{(i)} - \theta^T x^{(i)})^2$$

$$= -\frac{1}{2} \sum_{i=1}^{n} (y^{(i)} - \theta^T x^{(i)})^2$$

3. Solve

MLE – Linear Regression Model

$$p(y^{(i)}|x^{(i)};\theta) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(y^{(i)}-\theta^T x^{(i)})^2}{2\sigma^2}\right)$$

1. Determine formula for $LL(\theta)$

$$\log L(\theta) = n\log\frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{\sigma^2}\cdot\frac{1}{2}\sum_{i=1}^{n}(y^{(i)}-\theta^T x^{(i)})^2$$

2. Differentiate $LL(\theta)$ w.r.t. (each) θ

$$= -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y^{(i)}-\theta^T x^{(i)})^2$$

$$= -\frac{1}{2}\sum_{i=1}^{n}(y^{(i)}-\theta^T x^{(i)})^2$$

3. Solve

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, L(\theta)$$

$$= \operatorname{argmax}\left(-\frac{1}{2}\sum_{i=1}^{n}(y^{(i)}-\theta^T x^{(i)})^2\right)$$

$$= \operatorname{argmin}\left(\frac{1}{2}\sum_{i=1}^{n}(y^{(i)}-\theta^T x^{(i)})^2\right)$$

With probabilistic assumptions on the data, least-squares regression corresponds to finding the MLE of θ

MLE – Logistic Regression

$$y_i \mid x_i \sim \mathrm{Bern}(\sigma(\mathbf{w}^\top x_i))$$

1. Determine formula for $LL(\theta)$

$$LL(\theta) = \log \prod_{i=1}^{n} P_\theta\left(y^{(i)}|x^{(i)}\right) = \sum_{i=1}^{n} \log P_\theta\left(y^{(i)}|x^{(i)}\right)$$

$$p(y \mid x; \theta) = (h_\theta(x))^y \left(1 - h_\theta(x)\right)^{1-y}$$

2. Differentiate $LL(\theta)$ w.r.t. (each) θ

3. Solve

Mileage (in kmpl)	*Car Price (in cr)*
9.8	High
9.12	Low
9.5	High
10	Low

Bernoulli MLE Estimation

$X_1, X_2, \ldots, X_n$ where $X_i \sim \mathrm{Ber}(p)$. PMF of a Bernoulli $p^{X_i}(1-p)^{1-X_i}$

$$P_\theta(Y = 1|X = x) = h_\theta(x) = \frac{1}{1 + e^{-\theta^\top x}}$$

$$P_\theta(Y = 0|X = x) = 1 - h_\theta(x) = \frac{e^{-\theta^\top x}}{1 + e^{-\theta^\top x}}$$

MLE answers the question: For which parameter value does the observed data have the largest probability?

MLE – Logistic Regression

$$y_i \mid x_i \sim \text{Bern}(\sigma(\mathbf{w}^\top x_i))$$

1. Determine formula for $LL(\theta)$

$$L(\theta) = \prod_{i=1}^{n} P(Y = y^{(i)} | X = \mathbf{x}^{(i)})$$
$$= \prod_{i=1}^{n} \sigma(\theta^T \mathbf{x}^{(i)})^{y^{(i)}} \cdot \left[1 - \sigma(\theta^T \mathbf{x}^{(i)})\right]^{(1-y^{(i)})}$$

$$LL(\theta) = \sum_{i=1}^{n} y^{(i)} \log \sigma(\theta^T \mathbf{x}^{(i)}) + (1 - y^{(i)}) \log[1 - \sigma(\theta^T \mathbf{x}^{(i)})]$$

2. Differentiate $LL(\theta)$ w.r.t. (each) θ

$$\frac{\partial LL(\theta)}{\partial \theta_j} = \frac{\partial}{\partial \theta_j} y \log \sigma(\theta^T \mathbf{x}) + \frac{\partial}{\partial \theta_j} (1 - y) \log[1 - \sigma(\theta^T \mathbf{x}]$$
$$= \left[\frac{y}{\sigma(\theta^T \mathbf{x})} - \frac{1 - y}{1 - \sigma(\theta^T \mathbf{x})}\right] \frac{\partial}{\partial \theta_j} \sigma(\theta^T \mathbf{x})$$
$$= \left[\frac{y}{\sigma(\theta^T \mathbf{x})} - \frac{1 - y}{1 - \sigma(\theta^T \mathbf{x})}\right] \sigma(\theta^T \mathbf{x})[1 - \sigma(\theta^T \mathbf{x})]\mathbf{x}_j$$
$$= \left[\frac{y - \sigma(\theta^T \mathbf{x})}{\sigma(\theta^T \mathbf{x})[1 - \sigma(\theta^T \mathbf{x})]}\right] \sigma(\theta^T \mathbf{x})[1 - \sigma(\theta^T \mathbf{x})]\mathbf{x}_j$$
$$= [y - \sigma(\theta^T \mathbf{x})]\mathbf{x}_j$$

3. Solve

$$\theta := \theta + \alpha \nabla_\theta \ell(\theta)$$

MLE answers the question: For which parameter value does the observed data have the largest probability?

MLE – Logistic Regression

$$y_i \mid x_i \sim \mathrm{Bern}(\sigma(\mathbf{w}^\intercal x_i))$$

1. Determine formula for $LL(\theta)$

$$L(\theta) = \prod_{i=1}^{n} P(Y = y^{(i)} | X = \mathbf{x}^{(i)})$$
$$= \prod_{i=1}^{n} \sigma(\theta^T \mathbf{x}^{(i)})^{y^{(i)}} \cdot \left[1 - \sigma(\theta^T \mathbf{x}^{(i)})\right]^{(1-y^{(i)})}$$

$$LL(\theta) = \sum_{i=1}^{n} y^{(i)} \log \sigma(\theta^T \mathbf{x}^{(i)}) + (1 - y^{(i)}) \log[1 - \sigma(\theta^T \mathbf{x}^{(i)})]$$

2. Differentiate $LL(\theta)$ w.r.t. (each) θ

$$\frac{\partial LL(\theta)}{\partial \theta_j} = \sum_{i=1}^{n} \left[y^{(i)} - \sigma(\theta^T \mathbf{x}^{(i)})\right] x_j^{(i)}$$

3. Solve

MLE answers the question: For which parameter value does the observed data have the biggest probability?

MLE – Discrete - PMF

Example 1: Suppose that X is a discrete random variable with the following probability mass function: where $0 \leq \theta \leq 1$ is a parameter. The following 10 independent observations

X	0	1	2	3
$P(X)$	$2\theta/3$	$\theta/3$	$2(1-\theta)/3$	$(1-\theta)/3$

were taken from such a distribution: (3,0,2,1,3,2,1,0,2,1). What is the maximum likelihood estimate of θ.

1. Determine formula for $LL(\theta)$
2. Differentiate $LL(\theta)$ w.r.t. (each) θ
3. Solve

MLE – Discrete - PMF

1. Determine formula for $LL(\theta)$

   L(ɵ) = P(X=3)*P(X=0)*P(X=2)*……..*P(X=1)

   $= ((1-\theta) / 3)^2 * (2\theta / 3)^2 * (2(1-\theta) / 3)^3 * (\theta / 3)^3$

2. Differentiate $LL(\theta)$ w.r.t. (each) θ

3. Solve

X	P(X)
3	(1-ɵ) / 3
0	2ɵ / 3
2	2(1-ɵ) / 3
1	ɵ/ 3
3	(1-ɵ) / 3
2	2(1-ɵ) / 3
1	ɵ/ 3
0	2ɵ / 3
2	2(1-ɵ) / 3
1	ɵ/ 3

MLE – Discrete - PMF

1. Determine formula for $LL(\theta)$

L(ɵ) = P(X=3)*P(X=0)*P(X=2)*........*P(X=1)

$= ((1-\theta) / 3)^2 * (2\theta / 3)^2 * (2(1-\theta) / 3)^3 * (\theta / 3)^3$

2. Differentiate $LL(\theta)$ w.r.t. (each) θ

$LL(\theta) = \log [((1-\theta) / 3)^2 * (2\theta / 3)^2 * (2(1-\theta) / 3)^3 * (\theta/ 3)^3]$

$= \log ((1-\theta) / 3)^2 + \log (2\theta / 3)^2 + \log (2(1-\theta) / 3)^3 * \log(\theta/ 3)^3$

$= 2\log ((1-\theta) / 3) + 2\log (2\theta / 3) + 3 \log (2(1-\theta) / 3) + 3\log(\theta/ 3)$

$=2(\log ((1-\theta) -\log 3) + 2(\log (2\theta)-\log 3) + 3(\log (2(1-\theta)) - \log 3) + 3(\log(\theta)-\log 3)$

3. Solve

Gradient (LL(ɵ)) $= - \frac{2}{(1-\theta)} + \frac{2}{\theta} - \frac{3}{(1-\theta)} + \frac{3}{\theta}$

$= \frac{-5\theta + 5-5\theta}{(1-\theta)\theta} = \frac{-10\theta + 5}{(1-\theta)\theta} = 0 \rightarrow \theta = 0.5$

MLE-Summary

- Consider a sample of n i.i.d. random variables $X_1, X_2, \ldots, X_n$, drawn from a distribution $f(X_i|\theta)$
- θ_{MLE} maximizes the likelihood of data, $L(\theta)$ and $LL(\theta)$

$$L(\theta) = \prod_{i=1}^{n} f(X_i|\theta) \qquad \hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, L(\theta)$$

$$LL(\theta) = \log L(\theta) = \log \prod_{i=1}^{n} f(X_i|\theta) = \sum_{i=1}^{n} \log f(X_i|\theta)$$

$$\boxed{\theta_{MLE} = \arg\max_{\theta}\; LL(\theta)}$$

Maximum A Posteriori (MAP) Analysis

1. Determine prior probability

$$\theta_{MAP} = \arg\max f(\theta | X_1, X_2, \ldots, X_n)$$

2. Find the posterior probability for every distinct prior

   Brute Force MAP Hypothesis

$$h_{MAP} = \underset{h \in H}{argmax}\ P(h|D)$$
$$= \underset{h \in H}{argmax}\ \frac{P(D|h)P(h)}{P(D)}$$
$$= \underset{h \in H}{argmax}\ P(D|h)P(h)$$

3. Choose the posterior with highest h_{MAP} value

Maximum a Posteriori (MAP) Estimator of θ is the value of θ that maximizes the posterior distribution of θ.

Best hypothesis ≈ most probable hypothesis

Maximum A Posteriori Estimation (MAP)

1. Find the prior probability
2. Derive the posterior probability
3. Differentiate posterior w.r.t. (each) θ
4. Solve

Maximum a Posteriori (MAP) Estimator of θ is the value of θ that maximizes the posterior distribution of θ.

Best hypothesis ≈ most probable hypothesis

Example :

1. Example on MAP algorithm:

Let X be continuous random variable with probability density function P(X) given by:

$$f(x) = \begin{cases} 2x, & 0 \le x \le 1 \\ 0, & otherwise \end{cases}$$

Given another distribution $p(Y|X = x) = x(1-x)^{y-1}$ Find MAP estimate of X given Y=3

h_{MAP} = P(X | Y=3)

= P(Y=3 | X) * P(X)

= $x(1-x)^{y-1} * 2x$

To find the parameter X , differentiate the function & equate to zero.

$$\frac{d(P(X|Y=3)}{dx} = 0$$

$$\frac{d(x(1-x)2 * 2x\ )}{dx} = 0$$

$$\frac{d(2x^2-4x^3+2X^4)}{dx} = 0$$

$$4x-12x^2+8X^3 = 0$$

$$x = \{0, 0.5, 1\}$$

Example :

1. Example on MAP algorithm:

Let X be continuous random variable with probability density function P(X) given by:

$$f(x) = \begin{cases} 2x, & 0 \leq x \leq 1 \\ 0, & otherwise \end{cases}$$

Given another distribution $p(Y|X = x) = x(1-x)^{y-1}$ Find MAP estimate of X given Y=3

$$\begin{aligned} h_{MAP} &= P(X \mid Y=3) \\ &= P(Y=3 \mid X) * P(X) \\ &= x(1-x)^{y-1} * 2x \end{aligned}$$

x = {0, 0.5, 1}

P(X|Y=3) = {0, 0.125 ,0}

Most Probable Classification of New Instances

- So far we've sought the most probable hypothesis given the data D (i.e., h_{MAP})
- **Given new instance x, what is its most probable classification?**
  - $h_{MAP}(x)$ is not the most probable classification!
  - What's most probable classification of x?

Consider:

Three possible hypotheses:

$\boxed{P(h_1|D) = .4,}$ $P(h_2|D) = .3,$ $P(h_3|D) = .3$

Given new instance x, classification given by above 3 hypotheses is

$\boxed{h_1(x) = +,}$ $h_2(x) = -,$ $h_3(x) = -$

$P(\oplus|h_1) = 1$ and $P(\ominus|h_1) = 0$

May be the classification for X is **+**

Example 1 : Bayes Optimal Classifier

Bayes optimal classification: $\arg\max_{v_j \in V} \sum_{h_i \in H} P(v_j|h_i)P(h_i|D)$

- Example:

$P(h_1|D) = .4,\ P(-|h_1) = 0,\ P(+|h_1) = 1$

$P(h_2|D) = .3,\ P(-|h_2) = 1,\ P(+|h_2) = 0$

$P(h_3|D) = .3,\ P(-|h_3) = 1,\ P(+|h_3) = 0$

therefore

$$\sum_{h_i \in H} P(+|h_i)P(h_i|D) = .4$$

and

$$\sum_{h_i \in H} P(-|h_i)P(h_i|D) = .6$$

$$\boxed{\arg\max_{v_j \in V} \sum_{h_i \in H} P(v_j|h_i)P(h_i|D) = -}$$

Bayes' Optimal Classifier - Summary

- The most probable classification of the new instance is obtained by combining **the predictions of all hypotheses, weighted by their posterior probabilities.**
- v_j from some set V, then the probability $P(v_j \mid D)$ that the correct classification for the new instance is v_j is:

$$P(v_j|D) = \sum_{h_i \in H} P(v_j|h_i)P(h_i|D)$$

- The optimal classification of the new instance is the value vj for which $P(v_j \mid D)$ is maximum

BITS Pilani, Pilani Campus

Gibbs Classifier

- Bayes optimal classifier provides best result, but can be expensive if many hypotheses.
- Gibbs algorithm:
  - Choose one hypothesis at random, according to posterior prob. Distribution over h , $P(h|D)$
  - Use this h to classify new instance
- Surprising fact: under certain conditions, the expected misclassification error for the Gibbs algorithm is at most twice the expected error of the Bayes optimal classifier

$$E[error_{Gibbs}] \leq 2E\,[error_{BayesOptional}]$$

- Suppose correct, uniform prior distribution over H, then
  - Pick any hypothesis from *Version space*, with uniform probability
  - Its expected error no worse than twice Bayes optimal

Parameter Estimation in ML - Summary

Assumption: Data are__

- IID samples: $X_1 \ldots X_n$
  where all X_i are independent and have the same distribution.
- Either same PMF (discrete) or same PDF (continuous)
- $f(X \mid \theta)$
  Likelihood of different values of X depends on the values of our parameters θ
  f( ) is either PDF or PMF

$$\text{Parameters} : (\theta_0, \theta_1, \theta_2)$$

Find $\boldsymbol{\theta} = (\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ from the data $\mathbf{X_i}$

Maximum Likelihood Estimation (MLE)

select that parameters θ that make the observed data the most likely

$$f(X_1, X_2 \; \ldots.. X_n \mid \boldsymbol{\theta})$$

If the sample is large, MLE will yield an excellent estimator of θ
When no prior information is available, all hypothesis are equally likely i.e. p(hi) = p(hj)

Maximum A Posteriori (MAP)

choose the parameters θ that is the most likely, given the data

$$f(\boldsymbol{\theta} \mid X_1, X_2 \; \ldots.. X_n)$$

ML setting

- Bayesian Analysis
  - start with some belief about the system, called a prior.
  - Then we obtain some data and use it to update our belief.
  - The outcome is called a posterior.
  - Should we obtain even more data, the old posterior becomes a new prior and the cycle repeats.
  - P(h | D) a posterior determines the class label
  - MLE and MAP are the same if the prior is uniform
  - This forms the basis for Naïve Bayes classifier → Next Class

Previous Semester Exam from CANVAS shared papers
Answer Discussion

BITS Pilani

Maximum Likelihood Estimation (MLE)

1. Determine formula for $LL(\theta)$

2. Differentiate $LL(\theta)$ w.r.t. (each) θ

3. Solve

MLE – Example 1

Let $T_1, T_2, \ldots T_n$ be a random sample of a population describing the website loading time on a mobile browser with probability density function given as:

$$f(t/\theta) = \frac{1}{\theta} t^{\frac{(1-\theta)}{\theta}} \quad where \;\; 0 < t < 1 \; and \; 0 < \theta < \infty$$

Find the maximum likelihood estimator of θ. What is the estimate of θ, if the website loading time from four samples are $t_1 = 0.10$, $t_2 = 0.22$, $t_3 = 0.54$, $t_4 = 0.36$.

1. Determine formula for $LL(\theta)$
2. Differentiate $LL(\theta)$ w.r.t. (each) θ
3. Solve

t_i	$f(t \mid \theta)$
0.10	$\frac{1}{\theta} 0.10^{\frac{(1-\theta)}{\theta}}$
0.22	$\frac{1}{\theta} 0.22^{\frac{(1-\theta)}{\theta}}$
0.54	$\frac{1}{\theta} 0.54^{\frac{(1-\theta)}{\theta}}$
0.36	$\frac{1}{\theta} 0.36^{\frac{(1-\theta)}{\theta}}$

MLE – Example 1

1. Determine formula for $LL(\theta)$

$$L(\theta) = \frac{1}{\theta}\, 0.10^{\frac{(1-\theta)}{\theta}} * \frac{1}{\theta}\, 0.22^{\frac{(1-\theta)}{\theta}} * \frac{1}{\theta}\, 0.54^{\frac{(1-\theta)}{\theta}} * \ldots\ldots\ldots * \frac{1}{\theta}\, 0.36^{\frac{(1-\theta)}{\theta}}$$

$$= \theta^{-4}\, \left(\prod_{i=1}^{4} t_i\right)^{\frac{(1-\theta)}{\theta}}$$

2. Differentiate $LL(\theta)$ w.r.t. (each) θ

3. Solve

t_i	$f(t\|\theta)$
0.10	$\frac{1}{\theta}\, 0.10^{\frac{(1-\theta)}{\theta}}$
0.22	$\frac{1}{\theta}\, 0.22^{\frac{(1-\theta)}{\theta}}$
0.54	$\frac{1}{\theta}\, 0.54^{\frac{(1-\theta)}{\theta}}$
0.36	$\frac{1}{\theta}\, 0.36^{\frac{(1-\theta)}{\theta}}$

MLE – Example 1

1. Determine formula for $LL(\theta)$

   $L(\theta) = \theta^{-4} \left(\prod_{i=1}^{4} \mathbf{t}_\mathbf{i}\right)^{\frac{(1-\theta)}{\theta}}$

2. Differentiate $LL(\theta)$ w.r.t. (each) θ

   $LL(\theta) = \log \left[\theta^{-4} \left(\prod_{i=1}^{4} \mathbf{t}_\mathbf{i}\right)^{\frac{(1-\theta)}{\theta}}\right]$

   $= \log(\theta^{-4}) + \log \left(\prod_{i=1}^{4} \mathbf{t}_\mathbf{i}\right)^{\frac{(1-\theta)}{\theta}}$

   $= -4 \log(\theta) + \frac{(1-\theta)}{\theta} (\log(0.10 * 0.22 * 0.54 * 0.36))$

3. Solve

MLE – Example 1

1. Determine formula for $LL(\theta)$

$$L(\text{ɵ}) = \text{ɵ}^{-4}\left(\prod_{i=1}^{4}\mathbf{t}_{\mathbf{i}}\right)^{\frac{(1-\text{ɵ})}{\text{ɵ}}}$$

2. Differentiate $LL(\theta)$ w.r.t. (each) θ

$$LL(\text{ɵ}) = \log\left[\text{ɵ}^{-4}\left(\prod_{i=1}^{4}\mathbf{t}_{\mathbf{i}}\right)^{\frac{(1-\text{ɵ})}{\text{ɵ}}}\right]$$

$$= \log(\text{ɵ}^{-4}) + \log\left(\prod_{i=1}^{4}\mathbf{t}_{\mathbf{i}}\right)^{\frac{(1-\text{ɵ})}{\text{ɵ}}}$$

$$= -4\log(\text{ɵ}) + \frac{(1-\text{ɵ})}{\text{ɵ}}(\log(0.10*0.22*0.54*0.36))$$

$$= -4\log(\text{ɵ}) + \frac{1}{\text{ɵ}}(\log(0.10*0.22*0.54*0.36)) - (\log(0.10*0.22*0.54*0.36))$$

3. Solve

$$\text{Gradient}(LL(\text{ɵ})) = \frac{-4}{(\text{ɵ})} - \frac{(\log(0.004276))}{(\text{ɵ}^2)} = 0$$

$$\frac{-4}{(\text{ɵ})} = \frac{(\log(0.004276))}{(\text{ɵ}^2)}$$

$$\text{ɵ} = \frac{-(\log(0.004276))}{(4)}$$

ɵ = 1.3636 (base e) , ɵ = 1.9673 (base 2)

MLE – Example 2

Consider inputs x_i which are real valued attributes and the outputs y_i which are real valued of the form $y_i = f(x_i) + e_i$, where $f(x_i)$ is the true function and e_i is a random variable representing laplacian noise with PDF given by

$$\boxed{f\left(y_i/\theta\right) = \frac{1}{2\theta} * e^{\frac{-|y_i-\mu|}{\theta}}}$$

Implementing a linear regression model of the form , $h(x_i) = \sum_{i=0}^{n} \theta_i x_i$ and $\mu = h_i(x_i)$ find the maximum likelihood estimator of è. Comment on the loss function.

1. Determine formula for $LL(\theta)$
2. Differentiate $LL(\theta)$ w.r.t. (each) θ
3. Solve

MLE

1. Determine formula for $LL(\theta)$

$$L(\theta) = \frac{1}{2\theta} e^{\frac{-|y_1 - \mu|}{\theta}} * \frac{1}{2\theta} e^{\frac{-|y_2 - \mu|}{\theta}} * \frac{1}{2\theta} e^{\frac{-|y_3 - \mu|}{\theta}} * \ldots\ldots\ldots * \frac{1}{2\theta} e^{\frac{-|y_n - \mu|}{\theta}}$$

$$= \left(\prod_{i=1}^{n} \frac{1}{2\theta} e^{\frac{-|y_i - \mu|}{\theta}}\right)$$

2. Differentiate $LL(\theta)$ w.r.t. (each) θ

3. Solve

MLE

1. Determine formula for $LL(\theta)$

$$L(\theta) = \left(\prod_{i=1}^{n} \frac{1}{2\theta}\, e^{\frac{-|\mathbf{y_i} - \mu|}{\theta}}\right)$$

2. Differentiate $LL(\theta)$ w.r.t. (each) θ – natural log (ln)

$$LL(\theta) = \ln \left(\prod_{i=1}^{n} \frac{1}{2\theta}\, e^{\frac{-|\mathbf{y_i} - \mu|}{\theta}}\right)$$

$$= \sum_{i=1}^{n} \ln \left(\frac{1}{2\theta}\, e^{\frac{-|\mathbf{y_i} - \mu|}{\theta}}\right)$$

3. Solve

$$= -\ln(2\theta) \sum_{i=1}^{n} 1 + \sum_{i=1}^{n} \ln \left(e^{\frac{-|\mathbf{y_i} - \mu|}{\theta}}\right)$$

$$= \underset{\theta}{argmax} - n \ln(2\theta) - \sum_{i=1}^{n} \frac{|\mathbf{y_i} - \mu|}{\theta}$$

$$= \underset{\theta}{argmin}\; n \ln(2\theta) + \sum_{i=1}^{n} \frac{|\mathbf{y_i} - \mu|}{\theta}$$

MLE

1. Determine formula for $LL(\theta)$

   $L(\theta) = \left(\prod_{i=1}^{n} \frac{1}{2\theta} e^{\frac{-|\mathbf{y}_{\mathbf{i}} - \mu|}{\theta}}\right)$

2. Differentiate $LL(\theta)$ w.r.t. (each) θ

   $LL(\theta) = \ln \left(\prod_{i=1}^{n} \frac{1}{2\theta} e^{\frac{-|\mathbf{y}_{\mathbf{i}} - \mu|}{\theta}}\right)$

   $= \underset{\theta}{argmin}\ n \ln (2\theta) + \sum_{i=1}^{n} \frac{|\mathbf{y}_{\mathbf{i}} - \mu|}{\theta}$

3. Solve

   Gradient ($LL(\theta)$) $= \frac{n}{(\theta)} - \frac{\sum_{i=1}^{n} |\mathbf{y}_{\mathbf{i}} - \mu|}{(\theta^2)} = 0$

$$\frac{n}{(\theta)} = \frac{\sum_{i=1}^{n} |\mathbf{y}_{\mathbf{i}} - \mu|}{(\theta^2)}$$

$$\theta = \frac{\sum_{i=1}^{n} |\mathbf{y}_{\mathbf{i}} - \mu|}{\mathrm{n}}$$

Instead of MSE, MAE is the maximum likelihood hypothesis. So MAE is appropriate for the loss function

Additional References

- T1 book by Tom Mitchell – CH-6
- https://web.stanford.edu/class/archive/cs/cs109
- https://cs229.stanford.edu/lectures-spring2022/main_notes.pdf
- https://www.cs.cmu.edu/~ninamf/courses/601sp15/lectures.shtml

Thank you !

Required Reading for completed session :

T1 - Chapter # 6 (Tom M. Mitchell, Machine Learning)

R1 – Chapter # 3 (Christopher M. Bhisop, Pattern Recognition & Machine Learning)

& Refresh your MFDS & ISM parallel course basics

Next Session Plan :

Ensemble Learning & Naïve Bayes Classifier