Machine Learning

AIML CZG565

M8 : Bayesian Learning

BITS Pilani
Pilani Campus

Course Faculty of M.Tech Cluster
BITS – CSIS - WILP

Machine Learning

Disclaimer and Acknowledgement

- These content of modules & context under topics are planned by the course owner Dr. Sugata, with grateful acknowledgement to many others who made their course materials freely available online
- We here by acknowledge all the contributors for their material and inputs.
- We have provided source information wherever necessary
- Students are requested to refer to the textbook w.r.t detailed content of the presentation deck shared over canvas
- We have reduced the slides from canvas and modified the content flow to suit the requirements of the course and for ease of class presentation

Slide Source / Preparation / Review:
From BITS Pilani WILP: Prof.Sugata, Prof.Chetana, , Prof.Monali, Prof.Rajavadhana

External: CS109 and CS229 Stanford lecture notes, Dr.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 – Planned to be discussed in CS 15 or CS 16
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

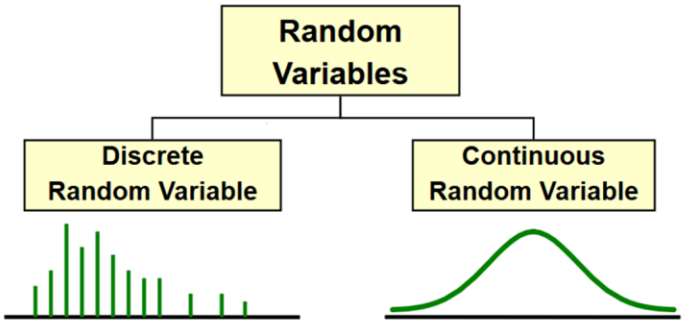
Mileage (in kmpl)	*Car Price (in cr)*
Neutral	High
Less	Low
Neutral	Medium
More	Low

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

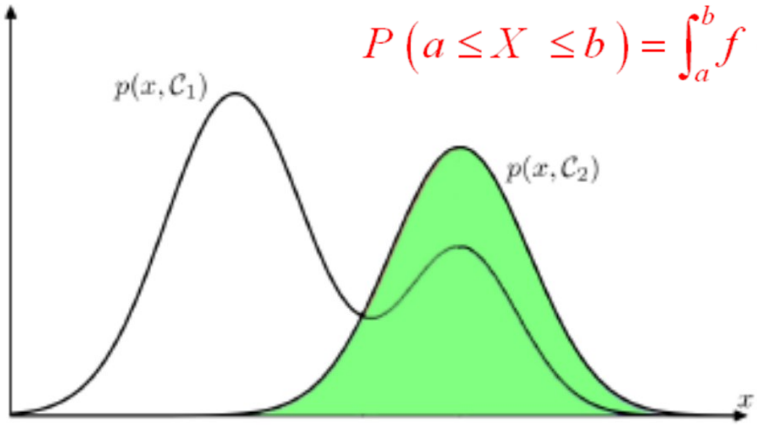
Represents a possible numerical value from a random event

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

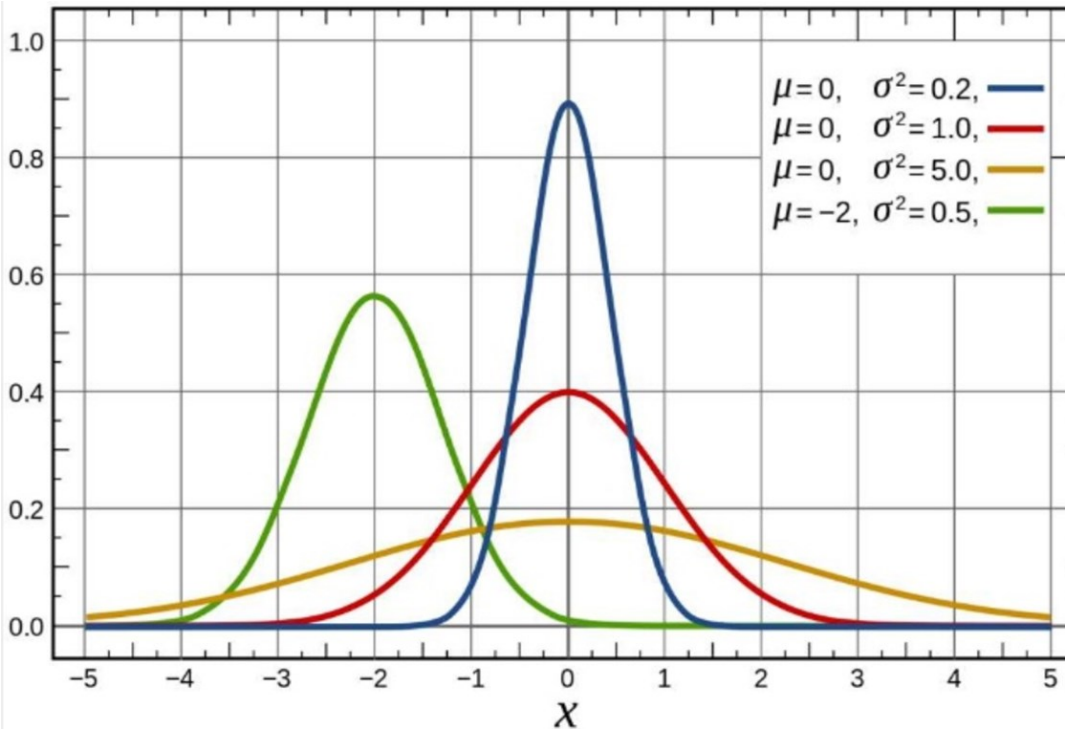
$$p(Y = y_j | X = x_i) = \frac{n_{ij}}{c_i}$$



$$P(a \leq X \leq b) = \int_a^b f(x)dx$$



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	$\theta = (m, b)$

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

θ is the parameter of a distribution.

θ can be a vector of parameters

Distribution = model + parameter θ

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 | θ)
  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 (θ | X_1 , X_2 .....X_n)

MLE: f (X_1 , X_2 .....X_n | θ)

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

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

$$\underbrace{P(\theta \mid \text{data})}_{\text{posterior}} = \frac{\overbrace{P(\text{data} \mid \theta)}^{\text{likelihood}}\overbrace{P(\theta)}^{\text{prior}}}{P(\text{data})}$$

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

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

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

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

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

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$_i$)

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

Maximum Likelihood Estimation (MLE)

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

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

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

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 $LL(\theta)$

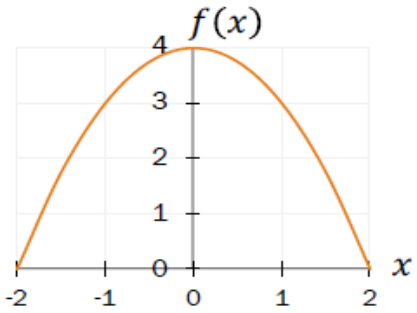
$$L(\theta) = \prod_{i=1}^{n} p(y^{(i)} \mid x^{(i)};\theta)$$

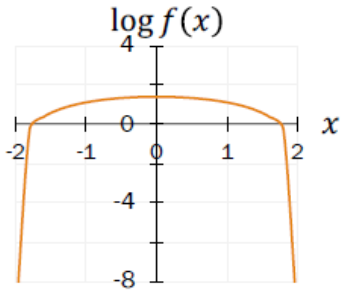
$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(y^{(i)}-\theta^T x^{(i)})^2}{2\sigma^2}\right)$$

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

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

3. Solve





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

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

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

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

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

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

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

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

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

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

MLE – 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

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

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

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

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

MLE – Logistic Regression

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

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

2. Differentiate $LL(\theta)$ w.r.t. (each) θ $\quad p(y \mid x; \theta) = (h_\theta(x))^y \left(1 - h_\theta(x)\right)^{1-y}$

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 $X_i \sim$ 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

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

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

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

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

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

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

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

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

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

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-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 $LL(\theta)$

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

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

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

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 $\approx$ 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.

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

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

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

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

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

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

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

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

x = {0, 0.5, 1}

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

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 → More on this will be discussed in later part of this module

Bayes' Optimal Classifier

- 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 I 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 I D) is maximum

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:

$P(h_1|D) = .4$, $P(h_2|D) = .3$, $P(h_3|D) = .3$

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

$h_1(x) = +$, $h_2(x) = -$, $h_3(x) = -$

$P(\oplus|h_1) = 1$ 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

and

$$\sum_{h_i \in H} P(+|h_i)P(h_i|D) = .4$$
$$\sum_{h_i \in H} P(-|h_i)P(h_i|D) = .6$$

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

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 | θ)
  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

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

Maximum A Posteriori (MAP)

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

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

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)

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

Additional Problems in MLE

BITS Pilani

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 $LL(\theta)$

$L(\theta) = P(X=3)*P(X=0)*P(X=2)*\ldots\ldots\ldots*P(X=1)$

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

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

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

Additional Exercise for Student's Practice

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

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

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

Naïve Bayes Classifier

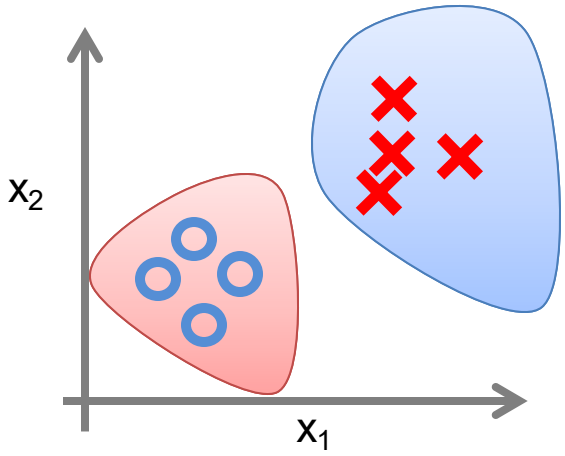
BITS Pilani

Refresh from Pre-Mid Term

Decision Theory: Interpretation

Model Building

Generative



$$P(Y \mid X_1 X_2 \ldots X_n) = \frac{P(X_1 X_2 \ldots X_d \mid Y) P(Y)}{P(X_1 X_2 \ldots X_d)}$$

Known as generative models, because by sampling from them it is possible to generate synthetic data points in the input space.

Eg., Gaussians, **Naïve Bayes**, Mixtures of multinomials **, Mixtures of Gaussians**, Bayesian networks

Likelihood — Class Prior Probability

$$P(c \mid x) = \frac{P(x \mid c) P(c)}{P(x)}$$

Posterior Probability — Predictor Prior Probability

$$P(c \mid \mathbf{X}) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

Sky	AirTemp	Humidity	Wind	Forecast	Enjoy Sport?
Sunny	Warm	Normal	Strong	Same	Yes
Sunny	Warm	High	Strong	Same	No
Rainy	Cold	High	Strong	Change	No
Sunny	Warm	Normal	Breeze	Same	Yes
Sunny	Hot	Normal	Breeze	Same	No
Rainy	Cold	High	Strong	Change	No
Sunny	Warm	High	Strong	Change	Yes
Rainy	Warm	Normal	Breeze	Same	Yes

Generative models for classification

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}.$$

- For binary classification the denominator is given by

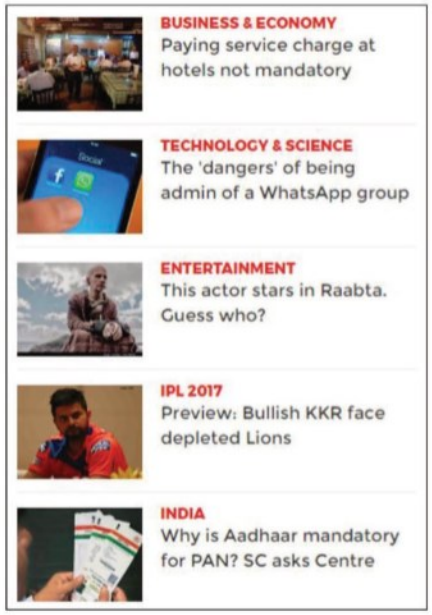
  $p(x) = p(x|y = 1)p(y = 1) + p(x|y = 0)p(y = 0)$

- if were calculating p(y|x) in order to make a prediction, then we don't actually need to calculate the denominator, since

$$\begin{aligned} \arg\max_y p(y|x) &= \arg\max_y \frac{p(x|y)p(y)}{p(x)} \\ &= \arg\max_y p(x|y)p(y). \end{aligned}$$

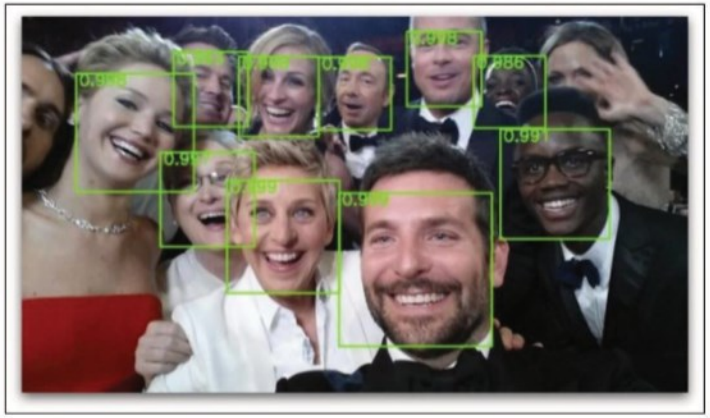
Naïve Bayes Classifier - Applications

Categorizing News



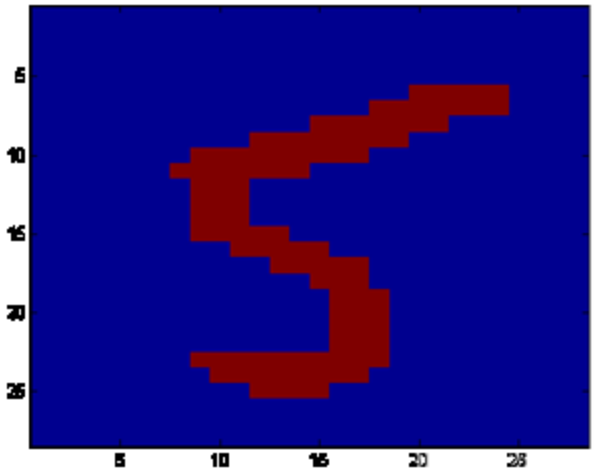
Email Spam Detection

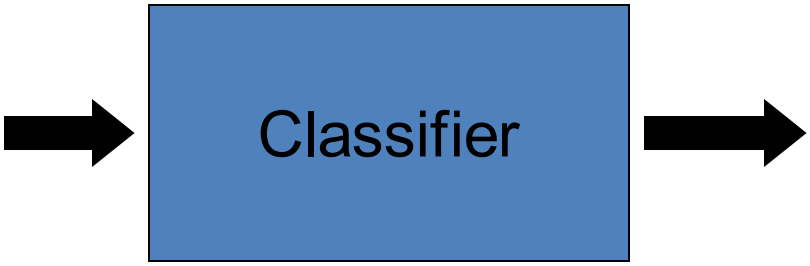
Face Recognition



Sentiment Analysis

Example: Digit Recognition





- $X_1,\ldots,X_n \in \{0,1\}$ (Black vs. White pixels)
- $Y \in \{5,6\}$ (predict whether a digit is a 5 or a 6)

The Bayes Classifier

$$P(Y=5|X_1,\ldots,X_n) = \frac{P(X_1,\ldots,X_n|Y=5)P(Y=5)}{P(X_1,\ldots,X_n|Y=5)P(Y=5)+P(X_1,\ldots,X_n|Y=6)P(Y=6)}$$

$$P(Y=6|X_1,\ldots,X_n) = \frac{P(X_1,\ldots,X_n|Y=6)P(Y=6)}{P(X_1,\ldots,X_n|Y=5)P(Y=5)+P(X_1,\ldots,X_n|Y=6)P(Y=6)}$$

- To classify, we'll simply compute these two probabilities and predict based on which one is greater

Naïve Bayes conditional Independence assumption

- Naïve Bayes assumes X_i are conditionally independent given Y

$$P(X_1|X_2, Y) = P(X_1|Y)$$

- **Assumption**:

$$P(X_1, \cdots, X_n|Y) = \prod_{j=1}^{n} P(X_j|Y)$$

i.e., X_i and X_j are conditionally independent given Y for $i \neq j$

Slide credit: Tom Mitchell

Naïve Bayes classifier: Prediction

Goal of learning P(Y|X) where X = <X_1,…, X_n>

- Bayes rule:

$$P(Y = y_k | X_1, \cdots, X_n) = \frac{P(Y = y_k) P(X_1, \cdots, X_n | Y = y_k)}{\sum_j P(Y = y_j) P(X_1, \cdots, X_n | Y = y_j)}$$

- Assume conditional independence among X_i's:

$$P(Y = y_k | X_1, \cdots, X_n) = \frac{P(Y = y_k) \Pi_i P(X_i | Y = y_k)}{\sum_j P(Y = y_j) \Pi_i P(X_i | Y = y_j)}$$

- Classify New Instance(*x*) : Pick the most probable (MAP) Y for

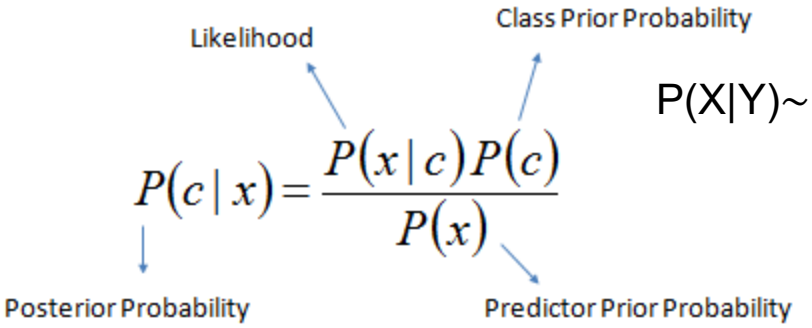
$$X_{new} = < X_1, \ldots, X_n >$$

$$\hat{Y} \leftarrow \underset{y_k}{\operatorname{argmax}} P(Y = y_k) \Pi_i P(X_i | Y = y_k)$$

Prior **Likelihood**

Slide credit: Tom Mitchell

Example: Play Tennis



$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

$$\hat{Y} \leftarrow \underset{y_k}{\operatorname{argmax}}\, P(Y = y_k) \Pi_i P(X_i | Y = y_k)$$

$P(X|Y) \sim \text{Multinom}(\pi, n) \rightarrow$ Multinomial NB (X_i – multinomial)

$P(Y) \sim$ Ber(p

Sky	AirTemp	Humidity	Wind	Forecast	Enjoy Sport?
Sunny	Warm	Normal	Strong	Same	Yes
Sunny	Warm	High	Strong	Same	No
Rainy	Cold	High	Strong	Change	No
Sunny	Warm	Normal	Breeze	Same	Yes
Sunny	Hot	Normal	Breeze	Same	No
Rainy	Cold	High	Strong	Change	No
Sunny	Warm	High	Strong	Change	Yes
Rainy	Warm	Normal	Breeze	Same	Yes

Example: Play Tennis – Learning Phase

Look up tables

Maximum likelihood estimates (MLE's):

$$\hat{\theta}_{ijk} = \hat{P}(X_i = x_{ij} | Y = y_k) = \frac{\#D\{X_i = x_{ij} \wedge Y = y_k\}}{\#D\{Y = y_k\}} \qquad \hat{\pi}_k = \hat{P}(Y = y_k) = \frac{\#D\{Y = y_k\}}{|D|}$$

Sky	Play=Yes	Play=No
Sunny	3/3	2/4
Rainy	0/3	2/4

Number of items in dataset D for which $Y=y_k$

Wind	Play=Yes	Play=No
Strong	2/3	3/4
Breeze	1/3	1/4

AirTemp	Play=Yes	Play=No
Hot	0/3	1/4
Warm	3/3	1/4
Cold	0/3	2/4

Forecast	Play=Yes	Play=No
Same	2/3	2/4
Change	1/3	2/4

Humidity	Play=Yes	Play=No
High	1/3	3/4
Normal	2/3	1/4

Sky	AirTemp	Humidity	Wind	Forecast	Enjoy Sport?
Sunny	Warm	Normal	Strong	Same	Yes
Sunny	Warm	High	Strong	Same	No
Rainy	Cold	High	Strong	Change	No
Sunny	Warm	Normal	Breeze	Same	Yes
Sunny	Hot	Normal	Breeze	Same	No
Rainy	Cold	High	Strong	Change	No
Sunny	Warm	High	Strong	Change	Yes
Rainy	Warm	Normal	Breeze	Same	Yes

Example: Play Tennis - Testing Phase

MAP rule

$$Y^{new} \leftarrow \arg\max_{y_k} \; P(Y = y_k) \prod_i P(X_i^{new} | Y = y_k)$$

$$Y^{new} \leftarrow \arg\max_{y_k} \; \pi_k \prod_i \theta_{ijk}$$

P(Enjoy=Yes | X) = P(X | Enjoy=Yes). P(Enjoy=Yes) / P(X)

= P(X | Enjoy=Yes). P(Enjoy=Yes)

= P(X | Enjoy=Yes). (3/7)

= P(Sunny | Enjoy=Yes). P(Warm | Enjoy=Yes). P(Normal | Enjoy=Yes). P(Strong | Enjoy=Yes). P(Change | Enjoy=Yes).(3/7)

= (3/3) . (3/3) . (2/3) . (2/3) . (1/3) . (3/7)

=0.0635

$$\hat{\pi}_k = \hat{P}(Y = y_k) = \frac{\#D\{Y = y_k\}}{|D|}$$

Sky	AirTemp	Humidity	Wind	Forecast	Enjoy Sport?
Sunny	Warm	Normal	Strong	Same	Yes
Sunny	Warm	High	Strong	Same	No
Rainy	Cold	High	Strong	Change	No
Sunny	Warm	Normal	Breeze	Same	Yes
Sunny	Hot	Normal	Breeze	Same	No
Rainy	Cold	High	Strong	Change	No
Sunny	Warm	High	Strong	Change	Yes
Sunny	Warm	Normal	Strong	Change	????
Rainy	Warm	Normal	Breeze	Same	????

P(Enjoy=Yes | X) > P(Enjoy=No | X) → EnjoySport = **Yes**

P(Enjoy=No | X) = P(X | Enjoy=No). P(Enjoy=No) / P(X)

= P(X | Enjoy=No). P(Enjoy=No)

= P(X | Enjoy=No). (4/7)

= P(Sunny | Enjoy=No). P(Warm | Enjoy=No). P(Normal | Enjoy=No). P(Strong | Enjoy=No). P(Change | Enjoy=No). (4/7)

= (2/4) . (1/4) . (1/4) . (3/4) . (2/4) . (4/7)

= 0.006696

Naïve Bayes Algorithm – discrete X_i

- Train Naïve Bayes (examples) for each* value y_k

  estimate $\pi_k \equiv P(Y = y_k)$

  for each* value x_{ij} of each attribute X_i

  estimate $\theta_{ijk} \equiv P(X_i = x_{ij} | Y = y_k)$

- Classify (X^{new})

$$Y^{new} \leftarrow \arg\max_{y_k} P(Y = y_k) \prod_i P(X_i^{new} | Y = y_k)$$

$$Y^{new} \leftarrow \arg\max_{y_k} \pi_k \prod_i \theta_{ijk}$$

Example: Play Tennis

P(Enjoy=Yes | X) = P(X | Enjoy=Yes). P(Enjoy=Yes) / P(X)

= P(X | Enjoy=Yes). P(Enjoy=Yes)

= P(X | Enjoy=Yes). (3/7)

= P(Rainy | Enjoy=Yes). P(Warm | Enjoy=Yes). P(Normal | Enjoy=Yes). P(Breeze | Enjoy=Yes). P(Same | Enjoy=Yes).(3/7)

= (0+1/3) . (3/3) . (2/3) . (1/3) . (2/3) . (3/7)

Sky	Enjoy Sport?
Sunny	Yes
Sunny	No
Rainy	No
Sunny	Yes
Sunny	No
Rainy	No
Sunny	Yes
Rainy	????

AirTemp	Humidity	Wind	Forecast	Enjoy Sport?
Warm	Normal	Strong	Same	Yes
Warm	High	Strong	Same	No
Cold	High	Strong	Change	No
Warm	Normal	Breeze	Same	Yes
Hot	Normal	Breeze	Same	No
Cold	High	Strong	Change	No
Warm	High	Strong	Change	Yes
Warm	Normal	Breeze	Same	????

P(Enjoy=No | X) = P(X | Enjoy=No). P(Enjoy=No) / P(X)

= P(X | Enjoy=No). P(Enjoy=No)

= P(X | Enjoy=No). (4/7)

= P(Rainy | Enjoy=No). P(Warm | Enjoy=No). P(Normal | Enjoy=No). P(Breeze | Enjoy=No). P(Same | Enjoy=No). (4/7)

= (2+1/4) . (1/4) . (1/4) . (1/4) . (2/4) . (4/7)

Naïve Bayes Classification Example

Name	Give Birth	Can Fly	Live in Water	Have Legs	Class
human	yes	no	no	yes	mammals
python	no	no	no	no	non-mammals
salmon	no	no	yes	no	non-mammals
whale	yes	no	yes	no	mammals
frog	no	no	sometimes	yes	non-mammals
komodo	no	no	no	yes	non-mammals
bat	yes	yes	no	yes	mammals
pigeon	no	yes	no	yes	non-mammals
cat	yes	no	no	yes	mammals
leopard shark	yes	no	yes	no	non-mammals
turtle	no	no	sometimes	yes	non-mammals
penguin	no	no	sometimes	yes	non-mammals
porcupine	yes	no	no	yes	mammals
eel	no	no	yes	no	non-mammals
salamander	no	no	sometimes	yes	non-mammals
gila monster	no	no	no	yes	non-mammals
platypus	no	no	no	yes	mammals
owl	no	yes	no	yes	non-mammals
dolphin	yes	no	yes	no	mammals
eagle	no	yes	no	yes	non-mammals

Give Birth	Can Fly	Live in Water	Have Legs	Class
yes	no	yes	no	?

A: attributes

M: mammals

N: non-mammals

$$P(A \mid M) = \frac{6}{7} \times \frac{6}{7} \times \frac{2}{7} \times \frac{2}{7} = 0.06$$

$$P(A \mid N) = \frac{1}{13} \times \frac{10}{13} \times \frac{3}{13} \times \frac{4}{13} = 0.0042$$

$$P(A \mid M)P(M) = 0.06 \times \frac{7}{20} = 0.021$$

$$P(A \mid N)P(N) = 0.004 \times \frac{13}{20} = 0.0027$$

P(A|M)P(M) > P(A|N)P(N)

=> Mammals

Issues with Naïve Bayes Classifier

Naïve Bayes Classifier:

P(Refund = Yes | No) = 3/7
P(Refund = No | No) = 4/7
P(Refund = Yes | Yes) = 0
P(Refund = No | Yes) = 1
P(Marital Status = Single | No) = 2/7
P(Marital Status = Divorced | No) = 1/7
P(Marital Status = Married | No) = 4/7
P(Marital Status = Single | Yes) = 2/3
P(Marital Status = Divorced | Yes) = 1/3
P(Marital Status = Married | Yes) = 0

For Taxable Income:
If class = No: sample mean = 110
sample variance = 2975
If class = Yes: sample mean = 90
sample variance = 25

- P(Yes) = 3/10
  P(No) = 7/10
- P(Yes | Married) = 0 x 3/10 / P(Married)
  P(No | Married) = 4/7 x 7/10 / P(Married)

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Source Credit : Slide adopted from "Introduction to Data mining" Vipin Kumar

Issues with Naïve Bayes Classifier

Consider the table with Tid = 7 deleted

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Naïve Bayes Classifier:

P(Refund = Yes | No) = 2/6
P(Refund = No | No) = 4/6
P(Refund = Yes | Yes) = 0
P(Refund = No | Yes) = 1
P(Marital Status = Single | No) = 2/6
P(Marital Status = Divorced | No) = 0
P(Marital Status = Married | No) = 4/6
P(Marital Status = Single | Yes) = 2/3
P(Marital Status = Divorced | Yes) = 1/3
P(Marital Status = Married | Yes) = 0/3
For Taxable Income:
If class = No: sample mean = 91
sample variance = 685
If class = No: sample mean = 90
sample variance = 25

Naïve Bayes will not be able to classify X as Yes or No!

Source Credit : Slide adopted from "Introduction to Data mining" Vipin Kumar

Laplace Smoothing

Smoothing

If one of the conditional probabilities is zero, then the entire expression becomes zero

- Technique for smoothing categorical data.
- A small-sample correction, or **pseudo-count**, will be incorporated in every probability estimate.
- No probability will be zero.

Smoothing

Probability estimation:

$$\text{Original}: P(A_i \mid C) = \frac{N_{ic}}{N_c}$$

$$\text{Laplace}: P(A_i \mid C) = \frac{N_{ic} + 1}{N_c + c}$$

$$\text{m-estimate}: P(A_i \mid C) = \frac{N_{ic} + mp}{N_c + m}$$

Bayesian approach

c: number of classes

N_c: number of instances in the class

N_{ic}: number of instances having attribute value A_i in class c

p: prior probability of the class

m: constant called the **equivalent sample size**, which determines how heavily to weight p relative to the observed data

Example: Play Tennis

P(Enjoy=Yes | X) = P(X | Enjoy=Yes). P(Enjoy=Yes) / P(X)

= P(X | Enjoy=Yes). P(Enjoy=Yes)

= P(X | Enjoy=Yes). (3/7)

= P(Rainy | Enjoy=Yes). P(Warm | Enjoy=Yes). P(Normal | Enjoy=Yes). P(Breeze | Enjoy=Yes). P(Same | Enjoy=Yes).(3/7)

= (0+1/3+2) . (3/3) . (2/3) . (1/3) . (2/3) . (3/7)

$$\hat{\theta}_i = \frac{x_i + \alpha}{N + \alpha d} \qquad (i = 1, \ \ldots, \ d),$$

Sky	Enjoy Sport?
Sunny	Yes
Sunny	No
Rainy	No
Sunny	Yes
Sunny	No
Rainy	No
Sunny	Yes
Rainy	????
Rainy	Yes
Sunny	Yes
Rainy	No
Sunny	No

AirTemp	Humidity	Wind	Forecast	Enjoy Sport?
Warm	Normal	Strong	Same	Yes
Warm	High	Strong	Same	No
Cold	High	Strong	Change	No
Warm	Normal	Breeze	Same	Yes
Hot	Normal	Breeze	Same	No
Cold	High	Strong	Change	No
Warm	High	Strong	Change	Yes
Warm	Normal	Breeze	Same	????

P(Enjoy=No | X) = P(X | Enjoy=No). P(Enjoy=No) / P(X)

= P(X | Enjoy=No). P(Enjoy=No)

= P(X | Enjoy=No). (4/7)

= P(Rainy | Enjoy=No). P(Warm | Enjoy=No). P(Normal | Breeze | Enjoy=No). P(Same | Enjoy=No). (4/7)

= (2+1/4+2) . (1/4) . (1/4) . (1/4) . (2/4) . (4/7)

Example: Play Tennis

P(Enjoy=Yes | X) = P(X | Enjoy=Yes). P(Enjoy=Yes) / P(X)

= P(X | Enjoy=Yes). P(Enjoy=Yes)

= P(X | Enjoy=Yes). (3/7)

= P(Rainy | Enjoy=Yes). P(Warm | Enjoy=Yes). P(Normal | Enjoy=Yes). P(Breeze | Enjoy=Yes). P(Same | Enjoy=Yes).(3/7)

= (1/5) . (3/3) . (2/3) . (1/3) . (2/3) . (3/7)

=0.0127

Sky	AirTemp	Humidity	Wind	Forecast	Enjoy Sport?
Sunny	Warm	Normal	Strong	Same	Yes
Sunny	Warm	High	Strong	Same	No
Rainy	Cold	High	Strong	Change	No
Sunny	Warm	Normal	Breeze	Same	Yes
Sunny	Hot	Normal	Breeze	Same	No
Rainy	Cold	High	Strong	Change	No
Sunny	Warm	High	Strong	Change	Yes
Sunny	Warm	Normal	Strong	Change	????
Rainy	Warm	Normal	Breeze	Same	????

P(Enjoy=Yes | X) > P(Enjoy=No | X) → EnjoySport = **Yes**

P(Enjoy=No | X) = P(X | Enjoy=No). P(Enjoy=No) / P(X)

= P(X | Enjoy=No). P(Enjoy=No)

= P(X | Enjoy=No). (4/7)

= P(Rainy | Enjoy=No). P(Warm | Enjoy=No). P(Normal | Enjoy=No). P(Breeze | Enjoy=No). P(Same | Enjoy=No). (4/7)

= (3/6) . (1/4) . (1/4) . (1/4) . (2/4) . (4/7)

= 0.0023

Naïve Bayes: Continuous Features

- X_i can be continuous

Naïve Bayes classifier:

$$Y = \arg\max_y P(Y = y) \prod_i P(X_i | Y = y)$$

Assumption: $P(X_i|Y)$ has a **Gaussian** distribution

The Gaussian Probability Distribution

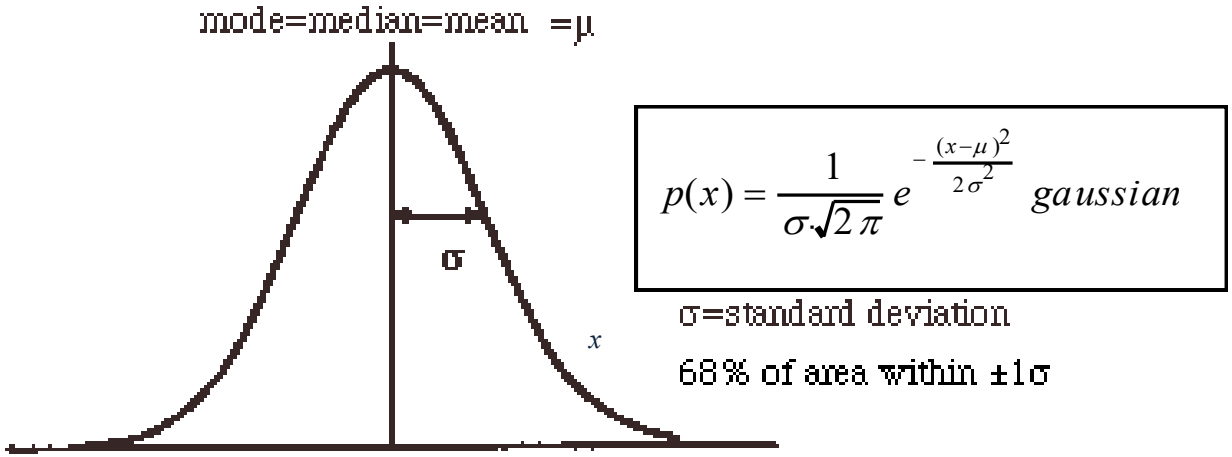
- It is a continuous distribution with pdf:

  μ = mean of distribution

  σ^2 = variance of distribution

  x is a continuous variable $(-\infty \leq x \leq)$

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad gaussian$$

σ=standard deviation

68% of area within ±1σ

Continuous Features : learning and prediction

- For each target value Y_k (MLE estimate)

$P(Y = y_k) \leftarrow$ No. of instances with Y_k class/No. of Total instances

- For each attribute value X_i estimate $P(X_i|Y = y_k)$
  - class conditional mean , variance

- Classify New Instance(x)

Pick the most probable (MAP) Y

$$\hat{Y} \leftarrow \underset{y_k}{\operatorname{argmax}} P(Y = y_k)\Pi_i P(X_i|Y = y_k)$$

Continuous Features : learning

- $P(X_i|Y)$ is Gaussian
- Training: estimate mean and standard deviation
  - $\mu_i = E[X_i|Y = y]$
  - $\sigma_i^2 = E[(X_i - \mu_i)^2|Y = y]$

X_1	X_2	X_3	Y
2	3	1	1
-1.2	2	0.4	1
1.2	0.3	0	0
2.2	1.1	0	1

Continuous Features : learning

X_1	X_2	X_3	Y
2	3	1	1
-1.2	2	0.4	1
1.2	0.3	0	0
2.2	1.1	0	1

- $\mu_i = E[X_i|Y = y]$
- $\sigma_i^2 = E[(X_i - \mu_i)^2|Y = y]$

- $\mu_1 = E[X_1|Y = 1] = \frac{2+(-1.2)+2.2}{3} = 1$
- $\sigma_1^2 = E[(X_1 - \mu_1)|Y = 1] =$

$$\frac{(2-1)^2 + (-1.2\ -1)^2 + (2.2\ -1)^2}{3} = 2.43$$

Estimating Parameters: X_i Continuous - Example

- E.g., character recognition: X_i is intensity at ith pixel

- Gaussian Naïve Bayes (GNB):

$$P(X_i = x|Y = y_k) = \frac{1}{\sigma_{ik}\sqrt{2\pi}} e^{-\frac{(x-\mu_{ik})^2}{2\sigma_{ik}^2}}$$

distribution of feature X_i is Gaussian with a mean and variance that can depend on the label y_k and which feature X_i it is

Estimating Parameters: X_i Continuous - Example

- E.g., character recognition: X_i is intensity at ith pixel

- Gaussian Naïve Bayes (GNB):

$$P(X_i = x | Y = y_k) = \frac{1}{\sigma_{ik}\sqrt{2\pi}} e^{-\frac{(x-\mu_{ik})^2}{2\sigma_{ik}^2}}$$

- Different mean and variance for each class k and each pixel i.
- Sometimes assume variance:
  - Is independent of Y (i.e., just have σ_i)
  - Or independent of X (i.e., just have σ_k)
  - Or both (i.e., just have σ)

Estimating Parameters: X_i Continuous - Example

- Maximum likelihood estimates:

$$\hat{\mu}_{MLE} = \frac{1}{N}\sum_{j=1}^{N} x_j$$

$$\hat{\mu}_{ik} = \frac{1}{\sum_j \delta(Y^j = y_k)} \sum_j X_i^j \delta(Y^j = y_k)$$

kth class

jth training image

ith pixel in jth training image

$$\hat{\sigma}^2_{unbiased} = \frac{1}{N-1}\sum_{j=1}^{N} (x_j - \hat{\mu})^2$$

$$\hat{\sigma}^2_{ik} = \frac{1}{\sum_j \delta(Y^j = y_k) - 1} \sum_j (X_i^j - \hat{\mu}_{ik})^2 \delta(Y^j = y_k)$$

Example: Play Tennis

$P(X|Y) \sim N(\mu, \sigma^2) \rightarrow$ GaussianNB (X_i – real valued)

P(Enjoy=Yes | X) = P(X | Enjoy=Yes). P(Enjoy=Yes) / P(X)

= P(X | Enjoy=Yes). P(Enjoy=Yes)

= P(X | Enjoy=Yes). (3/7)

= P(Rainy | Enjoy=Yes). P(Warm | Enjoy=Yes). P(60 | Enjoy=Yes). P(Breeze | Enjoy=Yes). P(Same | Enjoy=Yes).(3/7)

*= (1/3) . (3/3) . 0.15*10^{-95} . (1/3) . (2/3) . (3/7)*

$\mu_i = E[X_i|Y = yes] = 84.33$

$\sigma_i^2 = E[(X_i - \mu_i)^2|Y = yes] = 1.15$

$\mu_i = E[X_i|Y = no] = 72.5$

$\sigma_i^2 = E[(X_i - \mu_i)^2|Y = no] = 17.08$

Humidity	Enjoy Sport?
85	Yes
80	No
70	No
83	Yes
90	No
50	No
85	Yes
60	????

AirTemp	Sky	Wind	Forecast	Enjoy Sport?
Warm	Sunny	Strong	Same	Yes
Warm	Sunny	Strong	Same	No
Cold	Rainy	Strong	Change	No
Warm	Rainy	Breeze	Same	Yes
Hot	Sunny	Breeze	Same	No
Cold	Rainy	Strong	Change	No
Warm	Sunny	Strong	Change	Yes
Warm	Rainy	Breeze	Same	????

P(Enjoy=No | X) = P(X | Enjoy=No). P(Enjoy=No) / P(X)

= P(X | Enjoy=No). P(Enjoy=No)

= P(X | Enjoy=No). (4/7)

= P(Rainy | Enjoy=No). P(Warm | Enjoy=No). P(60 | Enjoy=No). P(Breeze | Enjoy=No). P(Same | Enjoy=No). (4/7)

= (2/4) . (1/4) . 0.02 . (1/4) . (2/4) . (4/7)

Additional Exercises – For Student's Practice

As a part of efforts to improve students' performance in the exams, you have been given the data showing number of study hours spent by students, their gender and their final results as pass or fail. Using this sample dataset, apply Naïve Bayes classification technique, to classify the below test case:

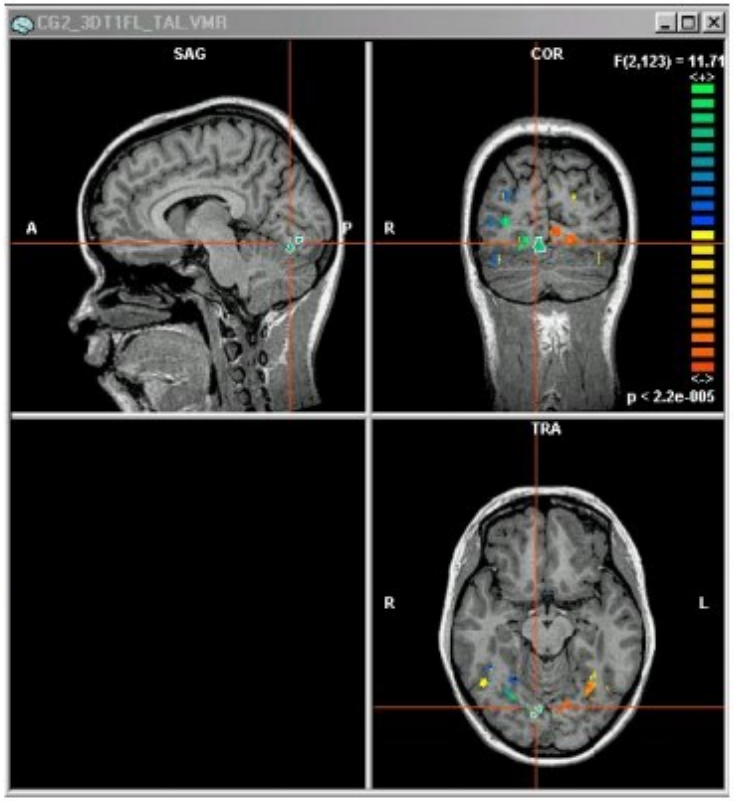
{No of study hours = 3.5, Gender="male"} either as "Pass", or "Fail".

Tip: $P(X|Y) \sim N(\mu, \sigma^2) \rightarrow$ GaussianNB (X_i – real valued)

No of study hours	Gender	Final result
4.5	Male	Pass
7	Female	Pass
2	Male	Fail
4	Female	Fail
2.5	Male	Fail
3	Female	Fail
8.3	Male	Fail
8	Female	Pass
9	Male	Pass

Example: GNB for classifying mental states

[Mitchell et al.]



- Classify a person's cognitive state, based on brain image
  - reading a sentence or viewing a picture?
  - reading the word describing a "Tool" or "Building"?
  - reading the word describing a "Person" or an "Animal"?
- Training: Patients were shown words of different categories and then a measurement was done to see what parts of the brain responded.

Example: GNB for classifying mental states

[Mitchell et al.]

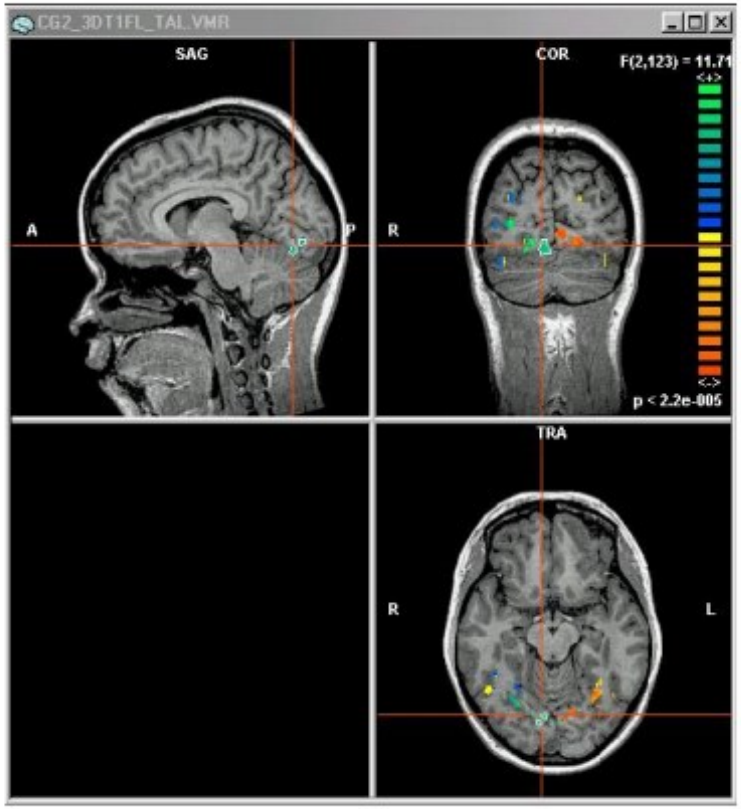
~1mm resolution

~2 images per sec.

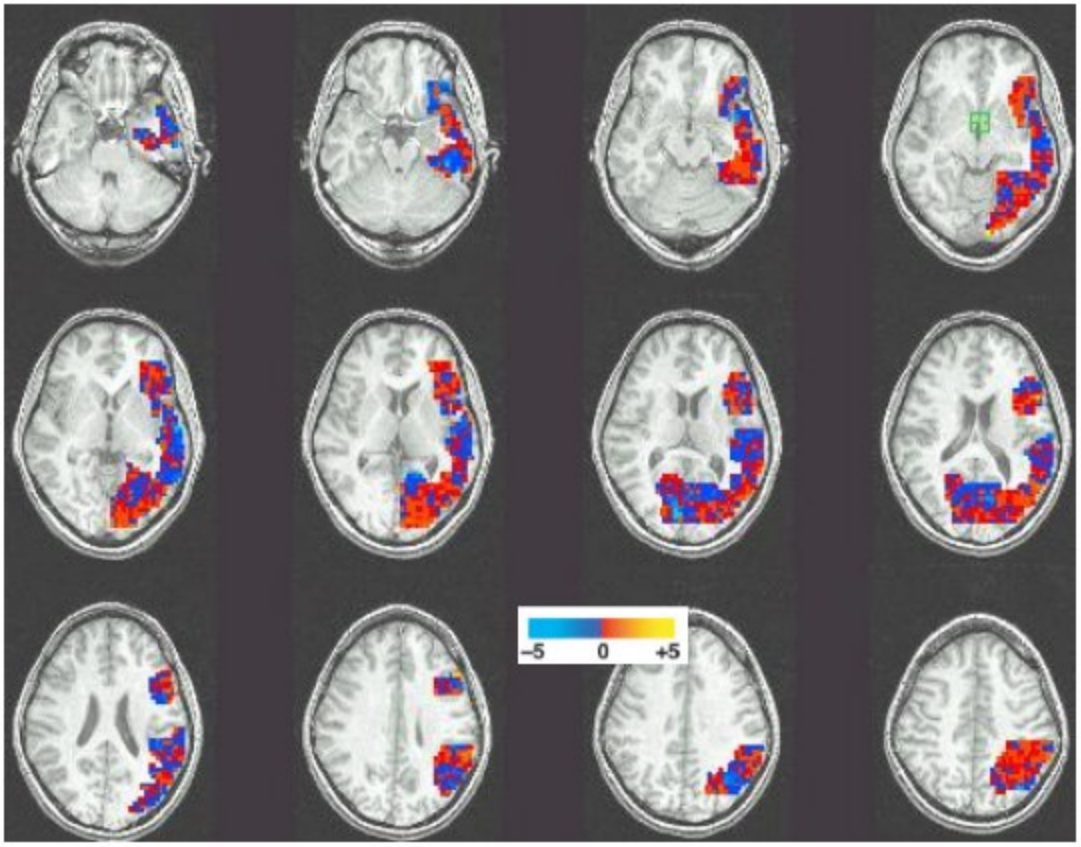
15,000 voxels/image

Non-invasive, save

Measures Blood Oxygen Level Dependent response (BOLD)



Gaussian Naïve Bayes: Learned $\mu_{voxel,word}$



[Mitchell et al.]

15,000 voxels or features

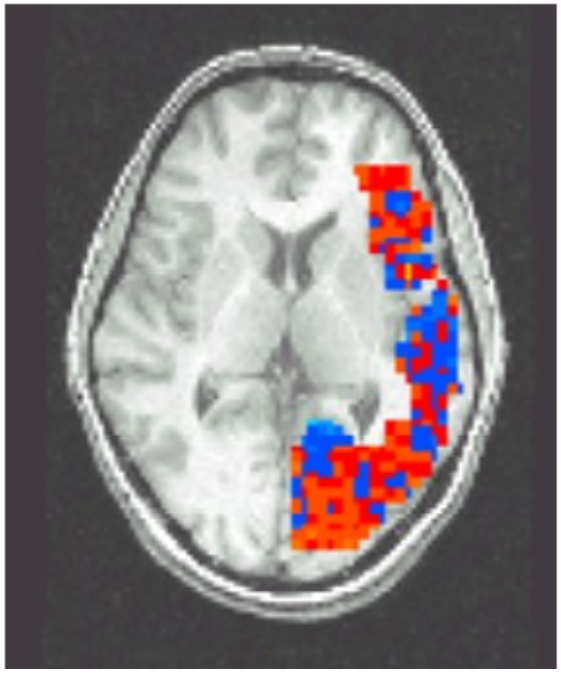
10 training examples or subjects per class

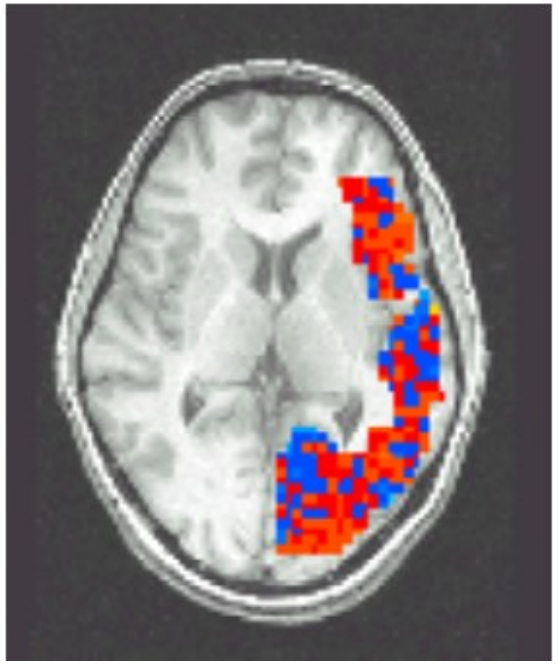
Learned Naïve Bayes Models –
Means for P(BrainActivity | WordCategory)

Pairwise classification accuracy: 85% [Mitchell et al.]

People words

Animal words





Text Classification using Naive Bayes Classifier

The Bag of Words Representation

I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!

it	6
I	5
the	4
to	3
and	3
seen	2
yet	1
would	1
whimsic al	1
times	1
sweet	1
satirical	1
adventur e	1
genre	1
fairy	1
humor	1
have	1
great	1

Example : Multinomial model :

Which Tag sentence " A very close game" belong to?

Text	Tag
"A great game"	Sports
"The election was over"	Not sports
"Very clean match"	Sports
"A clean but forgettable game"	Sports
"It was a close election"	Not sports

P(Sports=Yes) = 3/5
P(Sports=No) = 2/5
P(A | Sports=Yes) = 2/11
P(A | Sports=No) = 1/9

P(Sports=Yes | X) = P(X | Sports=Yes). P(Sports=Yes) / P(X)
= P(X | Sports=Yes). (3/5)
= P(A| Sports=Yes). P(Very | Sports=Yes). P(Close | Sports=Yes). P(Game | Sports=Yes). (3/5)
= (2/11) . (1/11) . (0/11) . (2/11) . (3/5)

A	Great	Game	The	Election	Was	Over	Very	Clean	Match	But	Forgettable	It	Close	*Sports or Not Sports*
1	1	1												1
			1	1	1	1								0
							1	1	1					1
1		1						1		1	1			1
1				1	1							1	1	0
1		1					1						1	????

Laplace Smoothing

- Laplace smoothing: we add 1 or in general constant k to every count so it's never zero.
- To balance this, we add the number of possible words to the divisor, so the division will never be greater than 1
- In our case, the possible words are ['a', 'great', 'very', 'over', 'it', 'but', 'game', 'election', 'clean', 'close', 'the', 'was', 'forgettable', 'match'].

- In our example
- we add 1 to every probability, therefore the probability, such as **P(close | sports)**, will never be 0.

Example : Multinomial model :

Which Tag sentence " A very close game" belong to?

Text	Tag
"A great game"	Sports
"The election was over"	Not sports
"Very clean match"	Sports
"A clean but forgettable game"	Sports
"It was a close election"	Not sports

P(Sports=Yes) = 3/5
P(Sports=No) = 2/5
P(A | Sports=Yes) = 2/11
P(A | Sports=No) = 1/9

P(Sports=Yes | X) = P(X | Sports=Yes). P(Sports=Yes) / P(X)
= P(X | Sports=Yes). (3/5)
= P(A| Sports=Yes). P(Very | Sports=Yes). P(Close | Sports=Yes). P(Game | Sports=Yes). (3/5)
= (2/11) . (1/11) . (0/11) . (2/11) . (3/5)
= (2+1/11+14) . (1+1/11+14) . (0+1/11+14) . (2+1/11+14) . (3/5)
= 0.00002765

(3+14/5+28)
=0.00002396

A	Great	Game	The	Election	Was	Over	Very	Clean	Match	But	Forgettable	It	Close	*Sports or Not Sports*
1	1	1												1
			1	1	1	1								0
							1	1	1					1
1		1						1		1	1			1
1				1	1							1	1	0
1		1					1						1	????

Apply Laplace Smoothing

Word	P(word | Sports)	P(word | Not Sports)
a	2+1 / 11+14	1+1 / 9+14
very	1+1 / 11+14	0+1 / 9+14
close	0+1 / 11+14	1+1 / 9+14
game	2+1 / 11+14	0+1 / 9+14

$$P(a|Sports) \times P(very|Sports) \times P(close|Sports) \times P(game|Sports) \times P(Sports)$$
$$= 2.76 \times 10^{-5}$$
$$= 0.0000276$$

$$P(a|Not\ Sports) \times P(very|Not\ Sports) \times P(close|Not\ Sports) \times P(game|Not\ Sports) \times P(Not\ Sports)$$
$$= 0.572 \times 10^{-5}$$
$$= 0.00000572$$

Summary: Learning to Classify Text

Target concept Interesting? : *Document* $\rightarrow\{+, -\}$

1. Represent each document by vector of words
   - one attribute per word position in document
2. Learning: Use training examples to estimate
   - $P(+)$
   - $P(-)$
   - $P(doc|+)$
   - $P(doc|-)$

Naive Bayes conditional independence assumption

$$P(doc|v_j) = \prod_{i=1}^{length(doc)} P(a_i = w_k|v_j)$$

where $P(a_i = w_k \mid v_j)$ is probability that word in position i is w_k, given v_j

one more assumption: $P(a_i = w_k|v_j) = P(a_m = w_k|v_j), \forall i, m$

Summary: Learning to Classify Text

LEARN_NAIVE_BAYES_TEXT (*Examples*, V)

1. *collect all words and other tokens that occur in Examples*

- *Vocabulary* ← all distinct words and other tokens in *Examples*

2. *calculate the required* $P(v_j)$ *and* $P(w_k \mid v_j)$ *probability terms*

- For each target value v_j in V do
  - $docs_j$ ← subset of *Examples* for which the target value is v_j
  - $Text_j$ ← a single document created by concatenating all members of $docs_j$

Summary: Learning to Classify Text

- $n \leftarrow$ total number of words in $Text_j$ (counting duplicate words multiple times)
- for each word w_k in *Vocabulary*
  - $n_k \leftarrow$ number of times word w_k occurs in $Text_j$
  - $$P(w_k|v_j) \leftarrow \frac{n_k+1}{n+|Vocabulary|}$$

CLASSIFY_NAIVE_BAYES_TEXT (*Doc*)

- *positions* $\leftarrow$ all word positions in *Doc* that contain tokens found in *Vocabulary*
- Return v_{NB} where $$v_{NB} = \underset{v_j \in V}{\operatorname{argmax}} P(v_j) \prod_{i \in positions} P(a_i|v_j)$$

Example 2: Multinomial model

	docID	words in document	in c = *China*?
training set	1	Chinese Beijing Chinese	yes
	2	Chinese Chinese Shanghai	yes
	3	Chinese Macao	yes
	4	Tokyo Japan Chinese	no
test set	5	Chinese Chinese Chinese Tokyo Japan	?

$$\hat{P}(w_t \mid C_k) = \frac{n_k(w_t)}{\sum_{s=1}^{|V|} n_k(w_s)},$$

N_{yes} (W=Chinese) = 5, N_{No} (W=Chinese) = 1,

|v| = 6 = {Chinese, Beijing, Shanghai, Macao, Tokyo Japan}
No of features (words) in Yes class = 8
No of features (words) in No class = 3

Source Credit : https://nlp.stanford.edu/IR-book

Example 2

	docID	words in document	in c = *China*?
training set	1	Chinese Beijing Chinese	yes
	2	Chinese Chinese Shanghai	yes
	3	Chinese Macao	yes
	4	Tokyo Japan Chinese	no
test set	5	Chinese Chinese Chinese Tokyo Japan	?

Priors: $\hat{P}(c) = 3/4$ and $\hat{P}(\overline{c}) = 1/4$

$$\hat{P}(w_t \mid C_k) = \frac{n_k(w_t)}{\sum_{s=1}^{|V|} n_k(w_s)}$$

$$\begin{aligned}
\hat{P}(\text{CHINESE}|c) &= (5+1)/(8+6) = 6/14 = 3/7 \\
\hat{P}(\text{TOKYO}|c) = \hat{P}(\text{JAPAN}|c) &= (0+1)/(8+6) = 1/14 \\
\hat{P}(\text{CHINESE}|\overline{c}) &= (1+1)/(3+6) = 2/9 \\
\hat{P}(\text{TOKYO}|\overline{c}) = \hat{P}(\text{JAPAN}|\overline{c}) &= (1+1)/(3+6) = 2/9
\end{aligned}$$

Source Credit : https://nlp.stanford.edu/IR-book

Example 2

	docID	words in document	in $c = China$?
training set	1	Chinese Beijing Chinese	yes
	2	Chinese Chinese Shanghai	yes
	3	Chinese Macao	yes
	4	Tokyo Japan Chinese	no
test set	5	Chinese Chinese Chinese Tokyo Japan	?

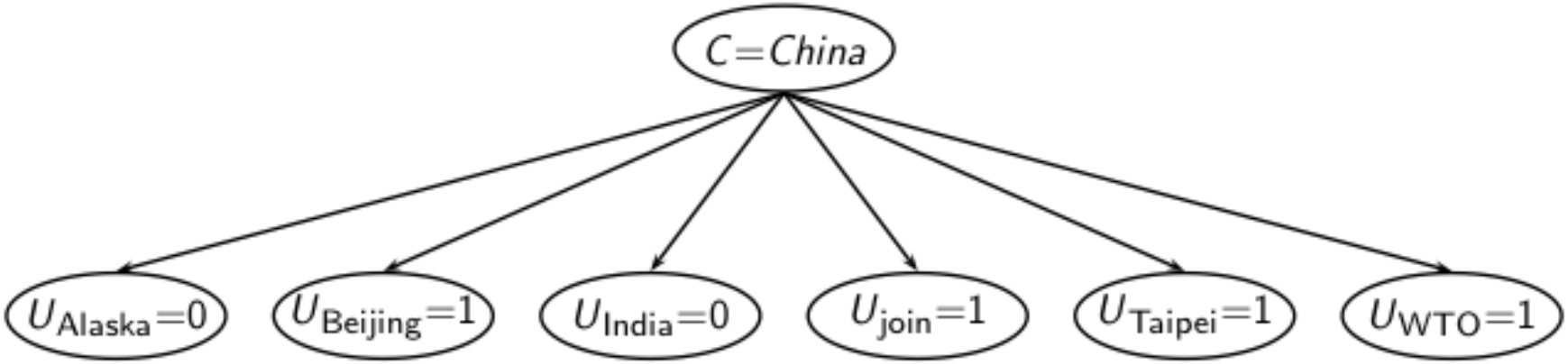
$$P(C_k|\mathcal{D}) \propto P(C_k) \prod_{j=1}^{\text{len}(\mathcal{D})} P(u_j|C_k)$$

u-each word in test document

$$\hat{P}(c|d_5) \propto 3/4 \cdot (3/7)^3 \cdot 1/14 \cdot 1/14 \approx 0.0003$$
$$\hat{P}(\overline{c}|d_5) \propto 1/4 \cdot (2/9)^3 \cdot 2/9 \cdot 2/9 \approx 0.0001$$

Source Credit : https://nlp.stanford.edu/IR-book

Different Naive Bayes model: Bernoulli model



One feature X_w for each word in dictionary

X_w = true in document d if w appears in d

Source Credit : https://nlp.stanford.edu/IR-book

Example 3

	docID	words in document	in c = *China*?
training set	1	Chinese Beijing Chinese	yes
	2	Chinese Chinese Shanghai	yes
	3	Chinese Macao	yes
	4	Tokyo Japan Chinese	no
test set	5	Chinese Chinese Chinese Tokyo Japan	?

$$\hat{P}(w_t \mid C_k) = \frac{n_k(w_t)}{N_k},$$

Let $n_k(w_t)$ be the number of documents of class k in which w_t is observed; and let N_k be the total number of documents of that class.

N_{yes} (W=Chinese) = 3, N_{No} (W=Chinese) = 1,

No of features (documents) in Yes class – (N_{Yes}) = 3

No of features (documents) in No class – (N_{No}) = 1

|v| = 6

Source Credit : https://nlp.stanford.edu/IR-book

Example 3

	docID	words in document	in c = *China*?
training set	1	Chinese Beijing Chinese	yes
	2	Chinese Chinese Shanghai	yes
	3	Chinese Macao	yes
	4	Tokyo Japan Chinese	no
test set	5	Chinese Chinese Chinese Tokyo Japan	?

$$\hat{P}(\text{Chinese}|c) = (3+1)/(3+2) = 4/5$$
$$\hat{P}(\text{Japan}|c) = \hat{P}(\text{Tokyo}|c) = (0+1)/(3+2) = 1/5$$
$$\hat{P}(\text{Beijing}|c) = \hat{P}(\text{Macao}|c) = \hat{P}(\text{Shanghai}|c) = (1+1)/(3+2) = 2/5$$
$$\hat{P}(\text{Chinese}|\overline{c}) = (1+1)/(1+2) = 2/3$$
$$\hat{P}(\text{Japan}|\overline{c}) = \hat{P}(\text{Tokyo}|\overline{c}) = (1+1)/(1+2) = 2/3$$
$$\hat{P}(\text{Beijing}|\overline{c}) = \hat{P}(\text{Macao}|\overline{c}) = \hat{P}(\text{Shanghai}|\overline{c}) = (0+1)/(1+2) = 1/3$$

Source Credit : https://nlp.stanford.edu/IR-book

Example 3

	docID	words in document	in c = *China*?
training set	1	Chinese Beijing Chinese	yes
	2	Chinese Chinese Shanghai	yes
	3	Chinese Macao	yes
	4	Tokyo Japan Chinese	no
test set	5	Chinese Chinese Chinese Tokyo Japan	?

b = feature vector for the document D

b_t, = {0,1} => absence or presence of word w_t in the document

$$P(C_k|\mathbf{b}) \propto P(\mathbf{b}|C_k)\,P(C_k)$$

$$\propto P(C_k)\prod_{t=1}^{|V|}\left[b_t\,P(w_t|C_k)+(1-b_t)\,(1-P(w_t|C_k))\right].$$

$$\begin{aligned}\hat{P}(c|d_5) \;\propto\;& \hat{P}(c)\cdot\hat{P}(\text{Chinese}|c)\cdot\hat{P}(\text{Japan}|c)\cdot\hat{P}(\text{Tokyo}|c)\\ &\cdot(1-\hat{P}(\text{Beijing}|c))\cdot(1-\hat{P}(\text{Shanghai}|c))\cdot(1-\hat{P}(\text{Macao}|c))\\ =\;& 3/4\cdot 4/5\cdot 1/5\cdot 1/5\cdot(1-2/5)\cdot(1-2/5)\cdot(1-2/5)\\ \approx\;& 0.005\end{aligned}$$

$$\begin{aligned}\hat{P}(\overline{c}|d_5) \;\propto\;& 1/4\cdot 2/3\cdot 2/3\cdot 2/3\cdot(1-1/3)\cdot(1-1/3)\cdot(1-1/3)\\ \approx\;& 0.022\end{aligned}$$

Source Credit : https://nlp.stanford.edu/IR-book

Naïve Bayes classifier: Summary Model

Model: joint probability distribution given by

- $P(X, Y) = P(Y)\ P(X|Y)$
- $P(X = X_1, \cdots, X_n, Y = y_k) = P(Y = y_k)\ P(X = X_1, \cdots, X_n | Y = y_k)$

Learning/Training:

For output variable Y

- P(Y) ~ Ber(p)

For each attribute X

- P(X|Y)~ $\text{Ber}(\pi)$ → Multivariate Bernoulli NB (X_i – binary)
- P(X|Y)~ $\text{Multinom}(\pi, n)$ → Multinomial NB (X_i – multinomial)
- P(X|Y) ~ N(μ, σ^2) → GaussianNB (X_i – real valued)

Source Credit : https://nlp.stanford.edu/IR-book

Logistic Regression vs Naïve Bayes

Idea:

- Naïve Bayes allows computing P(Y|X) by learning P(Y) and P(X|Y)

- Why not learn P(Y|X) directly?

Logistic Regression and Gaussian Naïve Bayes Classifier

- Interestingly, the parametric form of P(Y|X) used by Logistic Regression is precisely the form implied by the assumptions of a Gaussian Naive Bayes classifier.

- Therefore, we can view Logistic Regression as a closely related alternative to GNB, though the two can produce different results in many cases

- Derivation given in CS-5 slides in self learning section and also attached PDF of new chapter from Tom Mitchell book

Features of Bayesian learning

- Each observed training example can incrementally decrease or increase the estimated probability that a hypothesis is correct.

- Flexible approach to learning than algorithms that completely eliminate a hypothesis if it is found to be inconsistent with any single example.

Practical Issues of Bayesian learning

- Require initial knowledge of many probabilities
  - Often estimated based on background knowledge, previously available data, and assumptions about the form of the underlying distributions.
- Significant computational cost required to determine the Bayes optimal hypothesis in the general case (linear in the number of candidate hypotheses)

References

- https://www.inf.ed.ac.uk/teaching/courses/inf2b/learnnotes/inf2b-learn07-notes-nup.pdf
- https://cs229.stanford.edu/summer2019/cs229-notes2.pdf
- Tom Mitchell – Chapter 6
- https://nlp.stanford.edu/IR-book/pdf/13bayes.pdf

Logistic Regression & Naïve Bayes

Self Read

(Proof : Mathematical Relation between Logistic Regression and Naïve Bayes)

Where does the **form** come from?

- Logistic regression hypothesis representation

$$h_\theta(x) = \frac{1}{1 + e^{-\theta^\top x}} = \frac{1}{1 + e^{-(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n)}}$$

- Consider learning f: $X \rightarrow Y$, where
  - X is a vector of real-valued features $[X_1, \cdots, X_n]^\top$
  - Y is Boolean
  - Assume all X_i are conditionally independent given Y
  - Model $P(X_i | Y = y_k)$ as Gaussian $N(\mu_{ik}, \sigma_i)$
  - Model $P(Y)$ as Bernoulli π

What is $P(Y | X_1, X_2, \cdots, X_n)$?

Slide credit: Tom Mitchell

Where does the **form** come from?

- $P(Y=1|X) = \frac{P(Y=1)P(X|Y=1)}{P(Y=1)P(X|Y=1) + P(Y=0)P(X|Y=0)}$ Applying Bayes rule

$$= \frac{1}{1+\frac{P(Y=0)P(X|Y=0)}{P(Y=1)P(X|Y=1)}}$$ Divide by $P(Y=1)P(X|Y=1)$

$$= \frac{1}{1+\exp\left(\ln\left(\frac{P(Y=0)P(X|Y=0)}{P(Y=1)P(X|Y=1)}\right)\right)}$$ Apply $\exp(\ln(\cdot))$

$$= \frac{1}{1+\exp\left(\ln\left(\frac{1-\pi}{\pi}\right) + \sum_i ln\frac{P(X_i|Y=0)}{P(X_i|Y=1)}\right)}$$ Plug in $P(X_i|Y)$

$$P(x|y_k) = \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{(x-\mu_{ik})^2}{2\sigma_i^2}}$$

$$\sum_i \left(\frac{\mu_{i0}-\mu_{i1}}{\sigma_i^2} X_i + \frac{\mu_{i1}^2-\mu_{i0}^2}{2\sigma_i^2}\right)$$

$$P(Y=1|X_1, X_2, \cdots, X_n) = \frac{1}{1+\exp(\theta_0 + \sum_i \theta_i X_i)}$$

Slide credit: Tom Mitchell

Where does the **hypothesis function** come from?

- Logistic regression hypothesis representation

$$P(Y=1|X) = h_\theta(x) = \frac{1}{1+e^{-\theta^\top x}} = \frac{1}{1+e^{-(\theta_0+\theta_1 x_1+\theta_2 x_2+\cdots+\theta_n x_n)}}$$

  - Model **likelihood** $P(X_i|Y=y_k)$ as Gaussian $N(\mu_{ik},\sigma_i)$ and assume variance is independent of class, i.e. $\sigma_{i0} = \sigma_{i1} = \sigma_i$

$$P(x|y_k) = \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{(x-\mu_{ik})^2}{2\sigma_i^2}}$$

  - Model **prior** $P(Y)$ as Bernoulli π : **P(Y=1) = π and P(Y=0) = 1-π**

What is $P(Y|X_1, X_2, \cdots, X_n)$?

Slide credit: Tom Mitchell

Logistic Regression –Bayesian Analysis

$$P(Y=1|X) = \frac{P(Y=1)P(X|Y=1)}{P(Y=1)P(X|Y=1)+P(Y=0)P(X|Y=0)}$$

Applying Bayes rule

$$P(Y=1|X) = \frac{1}{1+\frac{P(Y=0)P(X|Y=0)}{P(Y=1)P(X|Y=1)}}$$

Divide by $P(Y=1)P(X|Y=1)$

$$P(Y=1|X) = \frac{1}{1+\exp(\ln\frac{P(Y=0)P(X|Y=0)}{P(Y=1)P(X|Y=1)})}$$

Apply $\exp(\ln(\cdot))$

$$P(Y=1|X) = \frac{1}{1+\exp(\ln\frac{P(Y=0)}{P(Y=1)}+\ln\frac{P(X|Y=0)}{P(X|Y=1)})}$$

Logistic Regression –Bayesian Analysis

By independence assumption:

$P(Y=1)=\pi$ and $P(Y=0)=1-\pi$
by modelling P(Y) as Bernoulli

$$\frac{P(X|Y=0)}{P(X|Y=1)} = \prod_i \frac{P(X_i|Y=0)}{P(X_i|Y=1)}$$

$$P(Y=1|X) = \frac{1}{1 + \exp\left(\ln\frac{1-\pi}{\pi} + \ln\prod_i \frac{P(X_i|Y=0)}{P(X_i|Y=1)}\right)}$$

$$P(Y=1|X) = \frac{1}{1 + \exp\left(\ln\frac{1-\pi}{\pi} + \sum_i \ln\frac{P(X_i|Y=0)}{P(X_i|Y=1)}\right)}$$

Logistic Regression –Bayesian Analysis

Plug in $P(X_i|Y)$

$$P(Y=1|X) = \frac{1}{1+\exp(\ln\frac{1-\pi}{\pi} + \sum_i \left(\frac{\mu_{i0}-\mu_{i1}}{\sigma_i^2}X_i + \frac{\mu_{i1}^2-\mu_{i0}^2}{2\sigma_i^2}\right))}$$

$$P(Y=1|X) = \frac{1}{1+\exp(w_0 + \sum_{i=1}^{n} w_i X_i)}$$

$$w_0 = \ln\frac{1-\pi}{\pi} + \sum_i \frac{\mu_{i1}^2-\mu_{i0}^2}{2\sigma_i^2}$$

$$w_i = \frac{\mu_{i0}-\mu_{i1}}{\sigma_i^2}$$

$$\begin{aligned}
\sum_i \ln\frac{P(X_i|Y=0)}{P(X_i|Y=1)} &= \sum_i \ln\frac{\frac{1}{\sqrt{2\pi\sigma_i^2}}\exp\left(\frac{-(X_i-\mu_{i0})^2}{2\sigma_i^2}\right)}{\frac{1}{\sqrt{2\pi\sigma_i^2}}\exp\left(\frac{-(X_i-\mu_{i1})^2}{2\sigma_i^2}\right)} \\
&= \sum_i \ln\exp\left(\frac{(X_i-\mu_{i1})^2-(X_i-\mu_{i0})^2}{2\sigma_i^2}\right) \\
&= \sum_i \left(\frac{(X_i-\mu_{i1})^2-(X_i-\mu_{i0})^2}{2\sigma_i^2}\right) \\
&= \sum_i \left(\frac{(X_i^2-2X_i\mu_{i1}+\mu_{i1}^2)-(X_i^2-2X_i\mu_{i0}+\mu_{i0}^2)}{2\sigma_i^2}\right) \\
&= \sum_i \left(\frac{2X_i(\mu_{i0}-\mu_{i1})+\mu_{i1}^2-\mu_{i0}^2}{2\sigma_i^2}\right) \\
&= \sum_i \left(\frac{\mu_{i0}-\mu_{i1}}{\sigma_i^2}X_i + \frac{\mu_{i1}^2-\mu_{i0}^2}{2\sigma_i^2}\right)
\end{aligned}$$

Features of Bayesian learning

- Each observed training example can incrementally decrease or increase the estimated probability that a hypothesis is correct.
- Flexible approach to learning than algorithms that completely eliminate a hypothesis if it is found to be inconsistent with any single example.
- Bayesian methods can accommodate hypotheses that make probabilistic predictions (e.g., hypotheses such as "this pneumonia patient has a 93% chance of complete recovery").

Features of Bayesian learning

- Prior knowledge can be combined with observed data to determine the final probability of a hypothesis.
- Prior knowledge is provided by asserting
  - prior probability for each candidate hypothesis, and
  - probability distribution over observed data for each possible hypothesis.
- New instances can be classified by combining the predictions of multiple hypotheses, weighted by their probabilities.

Practical Issues of Bayesian learning

- Require initial knowledge of many probabilities
  - Often estimated based on background knowledge, previously available data, and assumptions about the form of the underlying distributions.
- Significant computational cost required to determine the Bayes optimal hypothesis in the general case (linear in the number of candidate hypotheses)

Some Additional References

- T1 book by Tom Mitchell – Chapter-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
- https://nlp.stanford.edu/IR-book - Chapter 13

Thank you !

Required Reading for completed session :

T1 - Chapter # 8 (Tom M. Mitchell, Machine Learning)

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 :

Module 9 : Ensemble Learning