Bayes Decision Rule

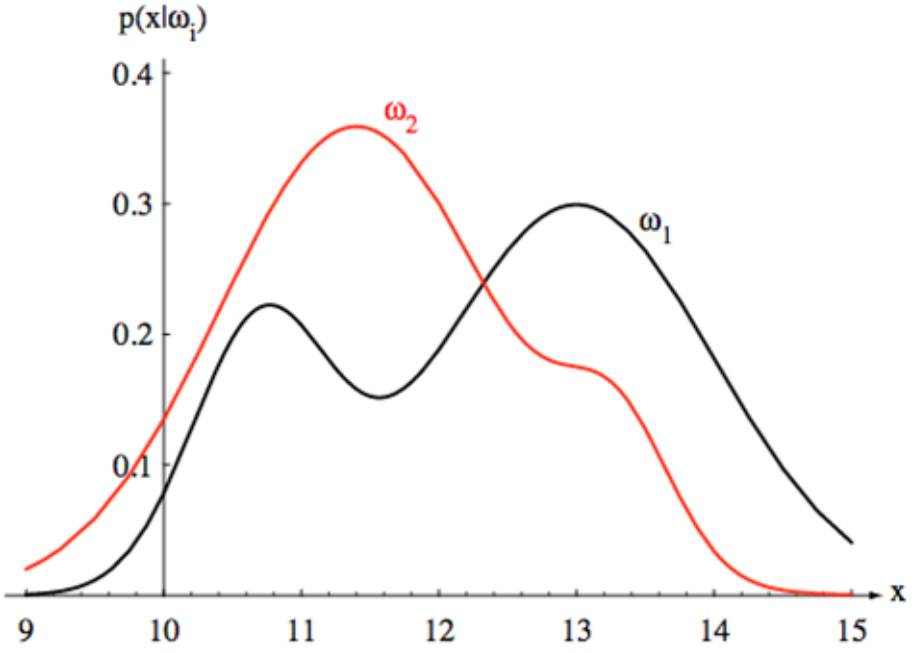
MLE, MAP, and Naïve Bayes

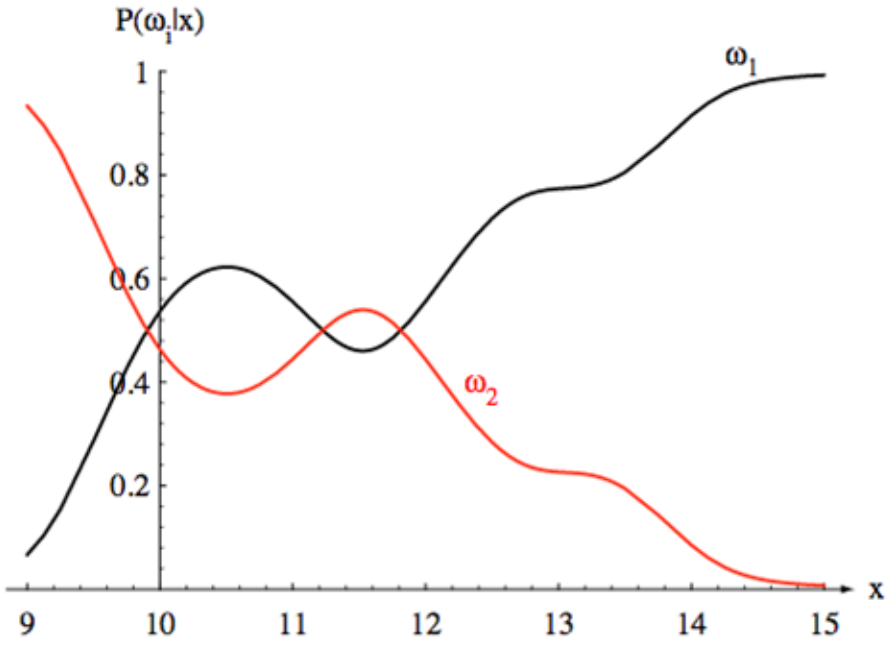
The Bayes Lecture

- Bayes Decision Rule
- Naïve Bayes

A simple decision rule

- If we can know either p(x|w) or p(w|x) we can make a classification guess

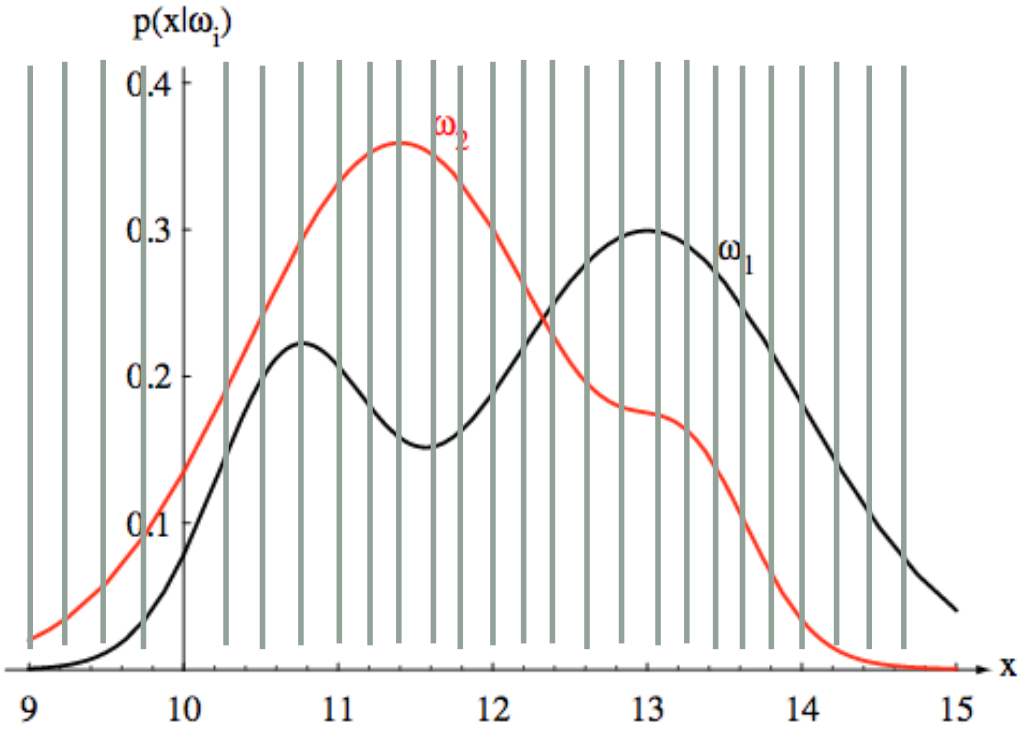




Goal: Find p(x|w) or p(w|x)

Duda et al. Pattern Classification

A simple way to estimate p(x|w)



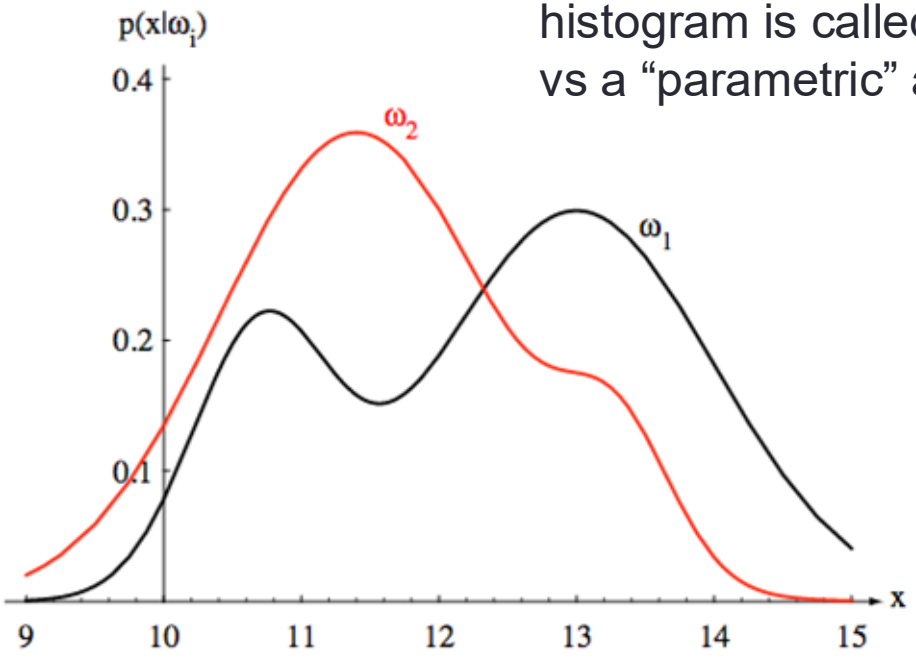
Make a histogram!

What happens if there is no data in a histogram bin?

The parametric approach

- We assume p(x|w) or p(w|x) follow some distributions with parameter θ

The method where we model the distribution using a histogram is called a “non-parametric” approach vs a “parametric” approach



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

Goal: Find θ so that we can estimate p(x|w) or p(w|x)

Maximum Likelihood Estimate (MLE)

$$p(w_i|x) = \frac{p(x|w_i)p(w_i)}{p(x)}$$

$$\text{Posterior} = \frac{\text{likelihood} * \text{prior}}{\text{evidence}}$$

- Maximizing the likelihood (probability of data given model parameters)

$p(\mathbf{x}|\theta) = L(\theta)$ <- This assumes the data is fixed

- Usually done on log likelihood

- Take the partial derivative wrt to θ and solve for the θ that maximizes the likelihood

MLE of Gaussian

- Observing $\{x_1, x_2, \ldots, x_i\}$, estimate the mean and the variance. Assume the data is normally distributed.

MLE of Gaussian $\theta = \mu$

- Observing $\{x_1, x_2, \ldots, x_n\}$, estimate the mean and the variance. Assume the data is normally distributed.

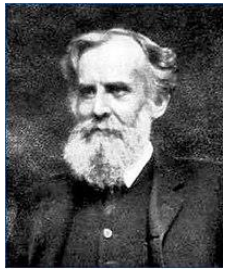
$$P(x_1, \ldots, x_n; \theta) \overset{iid}{=} \prod_{i=1}^{n} P(x_i; \theta)$$

$$\log \sum_{i=1}^{n} \log \left[\frac{1}{\sqrt{2\pi}\sigma^2} \exp\left(-\frac{1}{2} \frac{(x_i - \theta)^2}{\sigma^2} \right) \right]$$

$$= \sum_{i=1}^{n} \log \frac{1}{\sqrt{2\pi}\sigma^2} + \sum_{i=1}^{n} \left(-\frac{1}{2} \frac{(x_i - \theta)^2}{\sigma^2} \right) \quad (\log \to 0)$$

$$\frac{\partial}{\partial \theta} \sum_{i=1}^{n} \left[-\frac{1}{2} \cdot 2 \frac{(x_i - \theta)}{\sigma^2} (-1) \right] \to \theta = \frac{1}{n} \sum x_i$$

Frequentist vs Bayesian view



- Frequentist
  - Probability is “frequency of occurrence”
  - Data is from a random procedure that draw from unknown but fixed phenomenon.
    - Distribution parameter is a constant
- Bayesian
  - Probability is “degree of uncertainty”
  - Data is fixed and you want to infer about the unknown phenomenon.
    - Distribution parameter is a distribution
    - Prior knowledge about the phenomenon can change the inference results.

https://en.wikipedia.org/wiki/Frequentist_probability#/media/File:John_Venn.jpg
https://en.wikipedia.org/wiki/Thomas_Bayes#/media/File:Thomas_Bayes.gif

Maximum A Posteriori (MAP) Estimate

MLE

- Maximizing the likelihood (probability of data given model parameters)

$$\underset{\theta}{\text{argmax}}\ p(\mathbf{x}|\theta)$$

$$p(\mathbf{x}|\theta) = L(\theta)$$

- Usually done on log likelihood

- Take the partial derivative wrt to θ and solve for the θ that maximizes the likelihood

MAP

- Maximizing the posterior (model parameters given data)

$$\underset{\theta}{\text{argmax}}\ p(\theta|\mathbf{x})$$

- But we don't know $p(\theta|\mathbf{x})$

- Use Bayes rule

$$p(\theta|\mathbf{x}) = \frac{p(\mathbf{x}|\theta)p(\theta)}{p(\mathbf{x})}$$

- Taking the argmax for θ we can ignore $p(\mathbf{x})$

- $\underset{\theta}{\text{argmax}}\ p(\mathbf{x}|\theta)\ p(\theta)$

MAP on Gaussian

- We know x is Gaussian with an unknown mean μ that we need to estimate and a known variance σ^2
- Assume the prior of μ is $N(\mu_0, \sigma_0^2)$

- MAP estimate of μ is

$$\mu_n = \left(\frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2}\right)\left[\frac{1}{n}\Sigma_{i=1}^{n} x_i\right] + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2}\mu_0$$

$$\underset{\mu}{\operatorname{argmax}} \left[\prod_i P(x_i \mid \mu)\right] P(\mu)$$

$$= \prod_i \left[\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)\right] \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left(-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right)$$

$$\log \Rightarrow \sum \left[\log \frac{1}{\sqrt{2\pi}\sigma} - \frac{(x_i-\mu)^2}{2\sigma^2}\right] + \log \frac{1}{\sqrt{2\pi}\sigma_0} - \frac{(\mu-\mu_0)^2}{2\sigma_0^2}$$

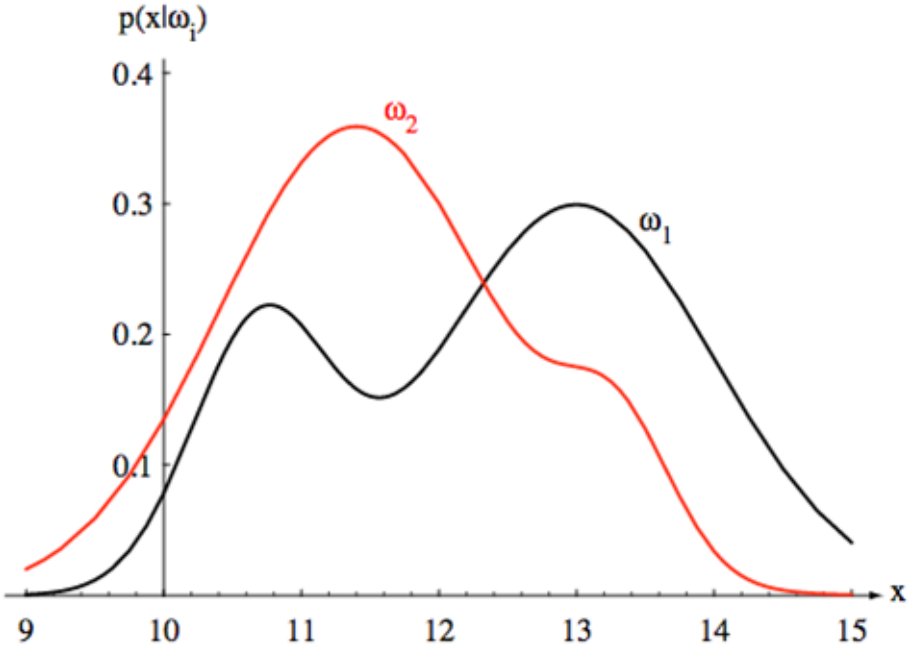
$$\frac{d}{d\mu} \quad - \sum^{n} \frac{2(x_i-\mu)(-1)}{2\sigma^2} - \frac{2(\mu-\mu_0)(-1)}{2\sigma_0^2}$$

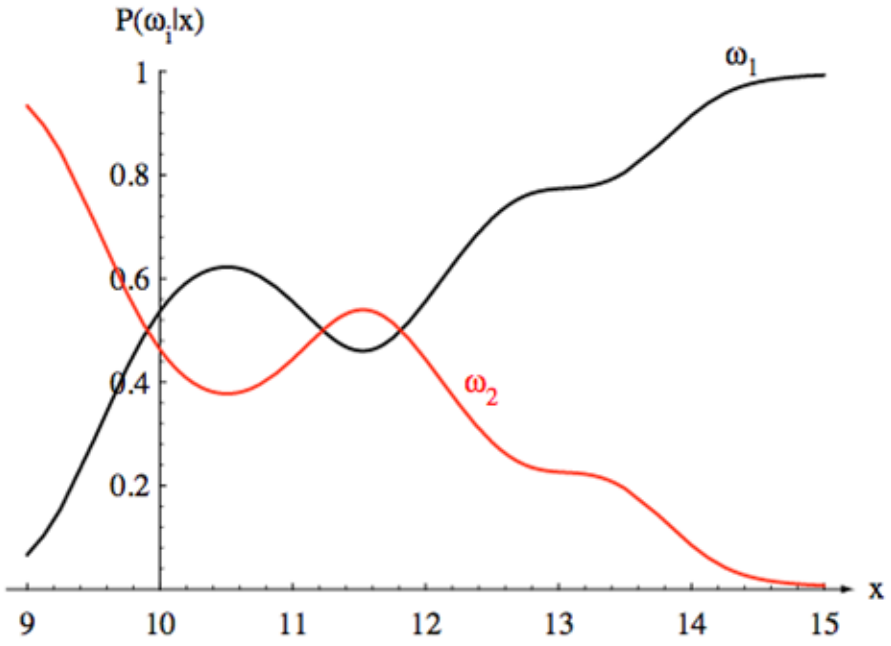
Notes of MAP estimate

- Usually harder to estimate than MLE
- If we use an uninformative prior for θ
  - MAP estimate = MLE
- Given infinite data
  - MAP estimate converges to MLE
- MAP is useful when you have less data, so you need additional knowledge about the domain
  - MAP estimate tends to converges to faster than MLE even with an arbitrary distribution
  - Can help prevent overfitting
- Useful for model adaptation (MAP adaptation)
  - Learn MLE on larger dataset, use this as your prior distribution
  - Learn MAP estimate on your dataset

A simple decision rule

- If we can know either p(x|w) or p(w|x) we can make a classification guess





Goal: Find p(x|w) or p(w|x) by finding the parameter of the distribution

Likelihood ratio test

- If $P(w_1|x) > P(w_2|x)$, that x is more likely to be class w_1
- Again we know $P(x|w_1)$ is more intuitive and easier to calculate than $P(w_1|x)$

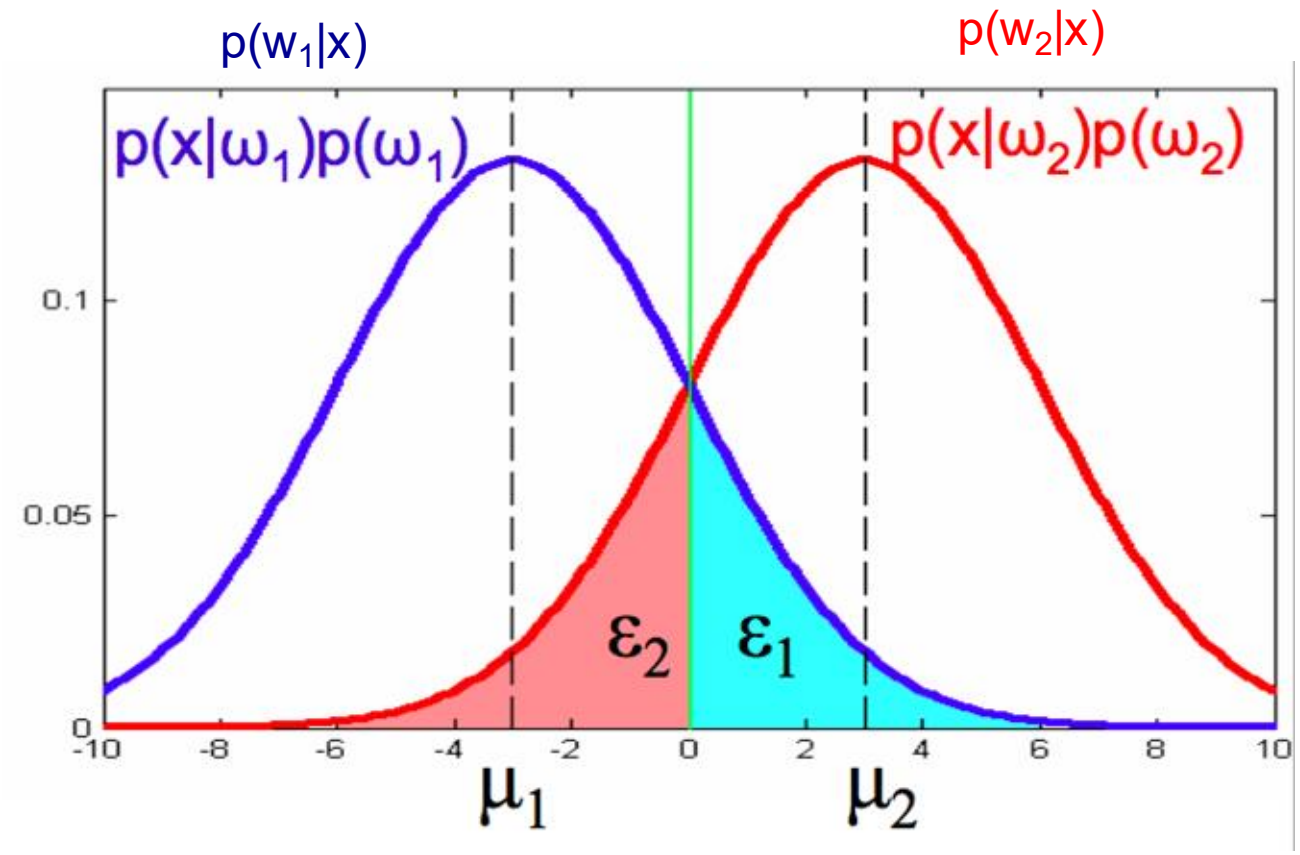
- Our classifier becomes
- $$P(x|w_1)P(w_1) \quad ? \quad P(x|w_2)P(w_2)$$
- $$\frac{P(x|w_1)}{P(x|w_2)} \quad ? \quad \frac{P(w_2)}{P(w_1)}$$

Likelihood ratio

Ratio of priors

Notes on likelihood ratio test (LRT)

- LRT minimizes the classification error (all errors are equally bad)

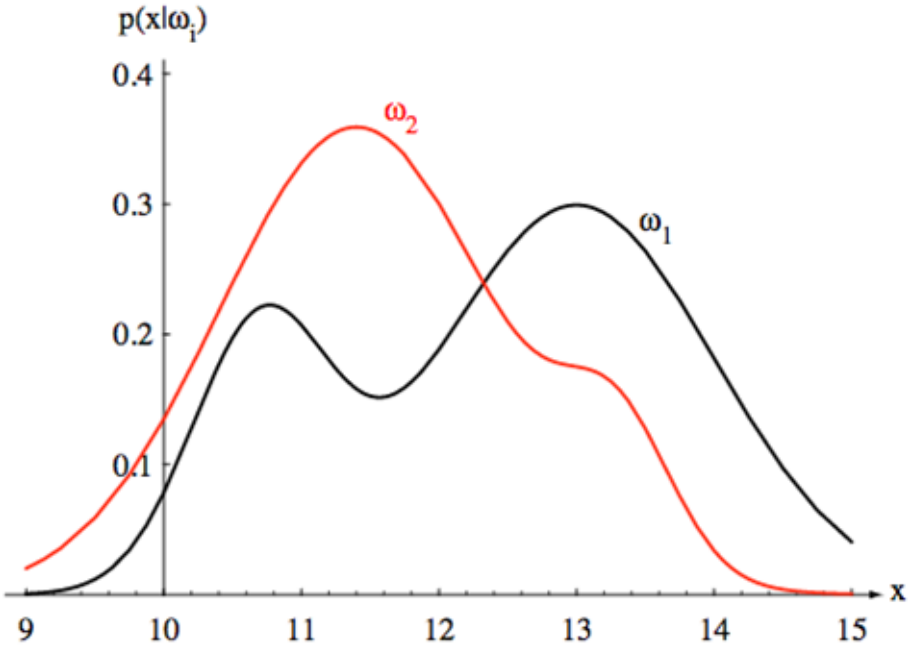


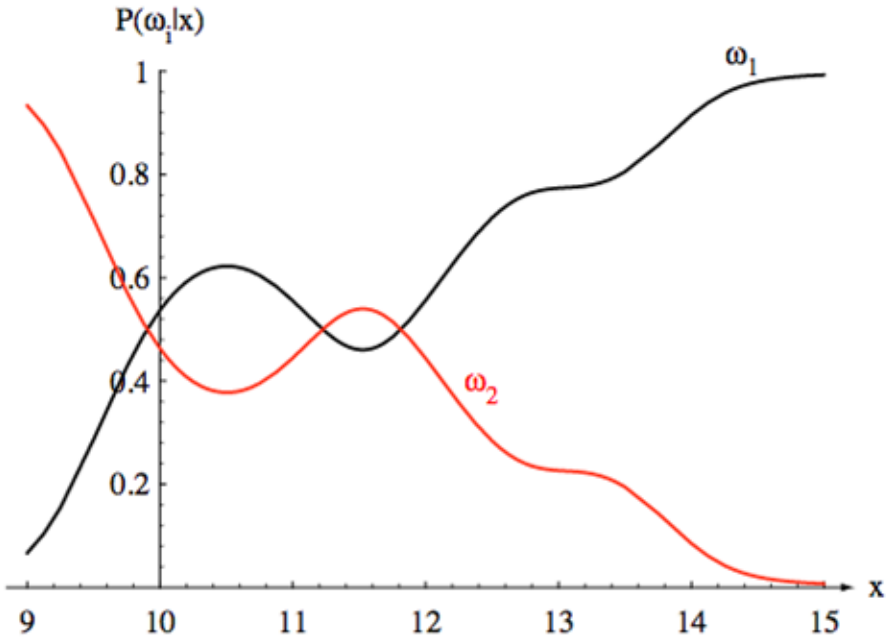
Notes on LRT

- If $P(w_1|x) > P(w_2|x)$, x is more likely to be class w_1
  - Also known as MAP decision rule
  - The classifier is sometimes called the Bayes classifier
- If we do not want to treat all error equally, we can assign different loss to each error, and minimize the expected loss. This is called Bayes loss/risk classifier
- $$\frac{P(x|w_1)}{P(x|w_2)} \quad ? \quad \frac{P(w_2)(L_{1|2} - L_{2|2})}{P(w_1)(L_{2|1} - L_{1|1})}$$
- When we treat errors equally, we refer to the zero-one loss
- $L_{1|2} = 1, L_{2|2} = 0, L_{2|1} = 1, L_{1|1} = 0$

Notes on LRT

- If we treat the priors as equal, we get the maximum likelihood criterion
- $$\frac{P(x|w_1)}{P(x|w_2)} \quad ? \quad 1$$





Naïve Bayes

- Below is the LRT or the Bayes classifier

$$P(x|w_1)P(w_1) \quad ? \quad P(x|w_2)P(w_2)$$

- What about Naïve Bayes?

- Here x is a vector with m features $[x_1, x_2, \ldots x_m]$
- $P(x|w_i)$ is m+1 dimensional
  - Sometimes to hard to model, not enough data, overfit, *curse of dimensionality*, etc.
- Assumes $x_1, x_2, \ldots x_m$ independent given w_i (conditional independence)
  - What does this mean?

Modeling distributions

Wind in the morning $X \in \{Calm, Windy\}$

PM2.5 level in the afternoon $Y \in \{Low, Med, High\}$

$\text{argmax } P(Y \mid X) = \text{argmax } P(X \mid Y)\, P(Y)$

Day	X	Y
1	W	M
2	C	M
3	W	M
4	W	H
5	C	L
6	W	L
7	C	H
8	W	L

Modeling distributions

Wind in the morning $X \in \{Calm, Windy\}$

PM2.5 level in the afternoon $Y \in \{Low, Med, High\}$

argmax P(Y | X)

Day	X	Y
1	W	M
2	C	M
3	W	M
4	W	H
5	C	L
6	W	L
7	C	H
8	W	L

P(X, Y)	L	M	H
C	1/8	1/8	1/8
W	2/8	2/8	1/8

Joint distribution

P(Y | X)	L	M	H
C	1/3	1/3	1/3
W	2/5	2/5	1/5

Conditional distribution

Modeling distributions

Wind in the morning $X \in \{Calm, Windy\}$

PM2.5 level in the afternoon $Y \in \{Low, Med, High\}$

argmax P(Y | X)

Day	X	Y
1	W	M
2	C	M
3	W	M
4	W	H
5	C	L
6	W	L
7	C	H
8	W	L

Joint distribution

P(X, Y)	L	M	H
C			
W			

Total data
8

count(X,Y)	L	M	H
C	1	1	1
W	2	2	1

$$P(X, Y) = \frac{\text{Count}(X, Y)}{\text{Total count}}$$ is the Maximum Likelihood Estimate (MLE) of P(X,Y)

Modeling distributions

Wind in the morning $X \in \{Calm, Windy\}$

PM2.5 level in the afternoon $Y \in \{Low, Med, High\}$

argmax P(Y | X)

Day	X	Y
1	W	M
2	C	M
3	W	M
4	W	H
5	C	L
6	W	L
7	C	H
8	W	L

P(Y | X)	L	M	H
C			
W			

count(X,Y)	L	M	H	Total
C	1	1	1	3
W	2	2	1	5

Conditional distribution

Total data
8

$$P(Y \mid X) = \frac{\text{Count}(X, Y)}{\text{Total count } (X)}$$ is the Maximum Likelihood Estimate (MLE) of P(Y|X)

Curse of dimensionality

Wind in the morning $X \in \{Calm, Windy\}$

PM2.5 level in the afternoon $Y \in \{Low, Med, High\}$

PM2.5 level in the evening $Z \in \{Low, Med, High\}$

$$\text{argmax } P(Z \mid Y, X) = \text{argmax } P(Y,X \mid Z)\, P(Z)$$

Day	X	Y	Z
1	W	L	M
2	C	M	M
3	W	H	M
4	W	M	H
5	C	M	L
6	W	M	L
7	C	L	H
8	W	H	L

count(Z,Y,X)	Z=L	Z=M	Z=H
X=W,Y=L	0	1	0
X=W,Y=M	1	0	1
X=W,Y=H	1	1	0
X=C,Y=L	0	0	1
X=C,Y=M	1	1	0
X=C,Y=H	0	0	0

Naïve Bayes

- $P(\mathbf{x}|w_i)P(w_i) = P(w_i) \prod_j P(x_j|w_i)$

- This assumption simplifies the calculation

Simplifying assumptions

Wind in the morning $X \in \{Calm, Windy\}$

PM2.5 level in the afternoon $Y \in \{Low, Med, High\}$

PM2.5 level in the evening $Z \in \{Low, Med, High\}$

$$\text{argmax } P(Z \mid Y, X) = \text{argmax } P(Y,X \mid Z)\, P(Z)$$
$$= \text{argmax } P(Y|Z)P(X|Z)P(Z)$$

Day	X	Y	Z
1	W	L	M
2	C	M	M
3	W	H	M
4	W	M	H
5	C	M	L
6	W	M	L
7	C	L	H
8	W	H	L

P(Y | Z)	Y = L	M	H
Z = L			
M			
H			

P(X | Z)	X = W	C
Z = L		
M		
H		

Dealing with zero probs

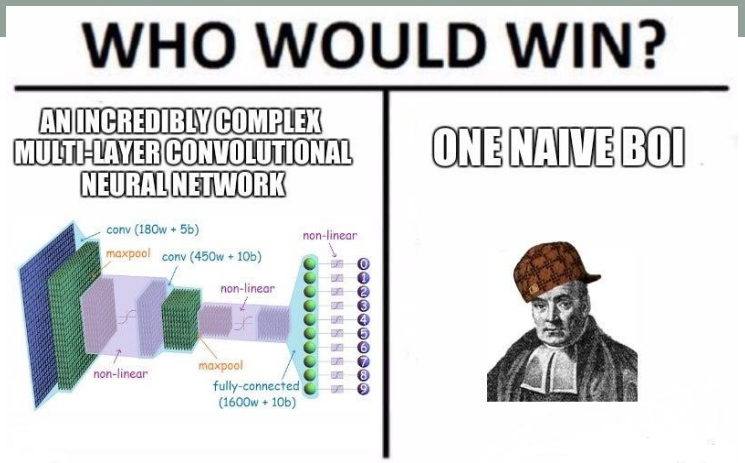
1. Use a very small value instead of zero (flooring)
2. Smooth the values using counts from other observations (smoothing)
3. Use priors (MAP adaptation)

Day	X	Y	Z
1	W	L	M
2	C	M	M
3	W	H	M
4	W	M	H
5	C	M	L
6	W	M	L
7	C	L	H
8	W	H	L

P(Y | Z)	Y = L	M	H
Z = L	0		
M			
H			

P(X | Z)	X = W	C
Z = L		
M		
H		

Naïve Bayes Notes



- $P(\mathbf{x}|w_i)P(w_i) = P(w_i) \prod_j P(x_j|w_i)$

- Note that we do not say anything about what kind of distribution $P(x_j|w_i)$ is.
  - In the homework you will play with this
    - Clean data
    - Estimate $P(x_j|w_i)$ using MLE, parametric and non-parametric version
    - Do prediction
    - Understand more about metrics
- Naïve Bayes can handle missing data
- Naïve Bayes is fast and quite good in practice
  - https://www.reddit.com/r/datascience/comments/hmhg9v/why_is_naive_bayes_so_popular_for_nlp/

Next homework

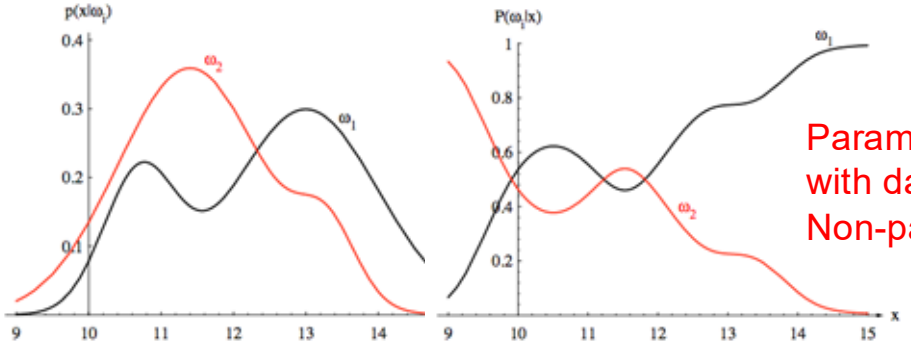
Summary

- MLE vs MAP estimate
  - How to use the prior
- LRT (Bayes Classifier)
  - Naïve Bayes

$$\frac{P(x|w_1)}{P(x|w_2)} \quad ? \quad \frac{P(w_2)}{P(w_1)}$$

Likelihood ratio

Ratio of priors



Parametric: need to assume the distribution (might not fit well with data)

Non-parametric: might encounter sparse bins