

# Bayesian classification

CISC 5800

Professor Daniel Leeds

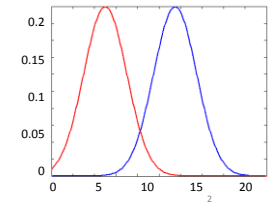
## Classifying with probabilities

Example goal: Determine is it cloudy out

- Available data: Light detector:  $x \in [0,25]$
- Potential class (atmospheric states):  $Y=\{\text{Cloudy}, \text{Non-Cloudy}\}$

Each class (atmospheric state)  $y$  has associated probability distribution  $P(x)$

Actually each  $y$  has a **likelihood** distribution  $P(x|\mu_y, \sigma_y)$

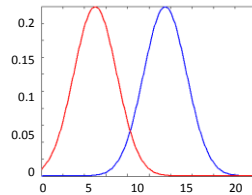


## Classifying with probabilities

Example goal: Determine is it cloudy out

- Measure light:  $x$
- Compute  $P(x|\mu_y, \sigma_y)$  for  $y=\text{Cloudy}$  and  $y=\text{Non-Cloudy}$
- Pick  $y$  which gives greatest **likelihood**  $P(x|\mu_y, \sigma_y)$   

$$\operatorname{argmax}_y P(x|\mu_y, \sigma_y)$$

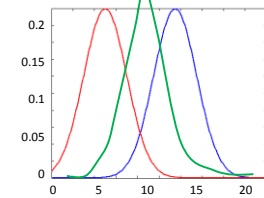


This is **Maximum Likelihood** classification

3

## What if there's an eclipse?

- Let's add a third potential class:  
 $Y=\{\text{Cloudy}, \text{Non-Cloudy}, \text{Eclipse}\}$
- What is most likely class if  $x=9$ ?
- Eclipses are low probability!



4

## Incorporating prior probability

- Define **prior** probabilities for each class  $P(y) = P(\mu_y, \sigma_y)$   
*Probability of class y same as probability of parameters  $\mu_y, \sigma_y$*
- **“Posterior probability”** estimated as likelihood  $\times$  prior :  
 $P(x|\mu_y, \sigma_y) P(\mu_y, \sigma_y)$
- Classify as  $\operatorname{argmax}_y P(x|\mu_y, \sigma_y) P(\mu_y, \sigma_y)$
- Terminology:  $\mu_y, \sigma_y$  are “parameters.” In general use  $\theta_y$   
Here:  $\theta_y = \{\mu_y, \sigma_y\}$ . **“Posterior”** estimate is  $P(x|\theta_y) P(\theta_y)$

5

## Probability review: Bayes rule

Recall:  $P(A|B) = \frac{P(A,B)}{P(B)}$

and:  $P(A, B) = P(B|A)P(A)$

so:  $P(A|B) = \frac{P(B|A) P(A)}{P(B)}$

Equivalently:  $P(y|x) = P(\theta_y|x) = P(\theta_y|D) = \frac{P(D|\theta_y) P(\theta_y)}{P(D)}$

The true  
posterior



6

## The posterior estimate

- $\operatorname{argmax}_{\theta_y} P(\theta_y|\mathbf{D}) \propto P(\mathbf{D}|\theta_y)P(\theta_y)$
- Posterior  $\propto$  Likelihood  $\times$  Prior     $\propto$  - means proportional  
We “ignore” the  $P(\mathbf{D})$  denominator  
because  $\mathbf{D}$  stays same while comparing  
different classes ( $y$  represented by  $\theta_y$ )

9

## Typical classification approaches

- MLE – Maximum Likelihood: Determine parameters/class which maximize probability of the data

$$\operatorname{argmax}_{\theta_y} P(\mathbf{D}|\theta_y)$$

- MAP – Maximum A Posteriori: Determine parameters/class that has maximum probability

$$\operatorname{argmax}_{\theta_y} P(\theta_y|\mathbf{D})$$

10

## Bernoulli distribution – coin flips

We have three coins with known biases (favoring heads or tails)  
How can we determine our current coin?

Flip  $K$  times to see which bias it has

Data ( $\mathbf{D}$ ): {HHTH, TTHH, TTTT, HTTT} Bias ( $\theta_y$ ):  $p_y$ , probability of H for coin  $y$

$$P(\mathbf{D}|\theta_y) = p_y^{|\mathbf{H}|}(1 - p_y)^{|\mathbf{T}|} \quad |\mathbf{H}| - \# \text{ heads}, \quad |\mathbf{T}| - \# \text{ tails}$$

12

## Bernoulli distribution – reexamined

$$P(\mathbf{D}|\theta_y) = p_y^{|\mathbf{H}|}(1 - p_y)^{|\mathbf{T}|} \quad |\mathbf{H}| - \# \text{ heads}, \quad |\mathbf{T}| - \# \text{ tails}$$

More rigorously: in  $K$  trials,  $side_k = \begin{cases} 0 & \text{if tails on flip } k \\ 1 & \text{if heads on flip } k \end{cases}$

$$P(\mathbf{D}|\theta_y) = \prod_k p_y^{side_k}(1 - p_y)^{(1-side_k)}$$

13

## Optimization: finding the maximum likelihood parameter for a fixed class (fixed coin)

$$\operatorname{argmax}_{\theta} P(\mathbf{D}|\theta_y) = p_y - \text{probability of Head}$$

$$\operatorname{argmax}_p p_y^{|\mathbf{H}|}(1 - p_y)^{|\mathbf{T}|}$$

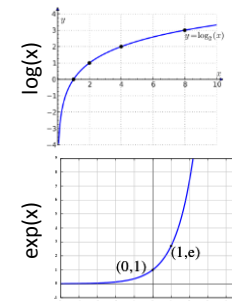
Equivalently, maximize  $\log P(\mathbf{D}|\theta_y)$

$$\operatorname{argmax}_{p_y} |\mathbf{H}| \log p_y + |\mathbf{T}| \log(1 - p_y)$$

15

## The properties of logarithms

- $e^a = b \leftrightarrow \log b = a$
- $a < b \leftrightarrow \log a < \log b$
- $\log ab = \log a + \log b$
- $\log a^n = n \log a$



Convenient when dealing with small probabilities

$$\bullet 0.0000454 \times 0.000912 = 0.0000000414 \rightarrow -10 + -7 = -17$$

16

## Intuition of the MLE result

$$p_y = \frac{|H|}{|H| + |T|}$$

- Probability of getting heads is # heads divided by # total flips

19

## Finding the maximum a posteriori

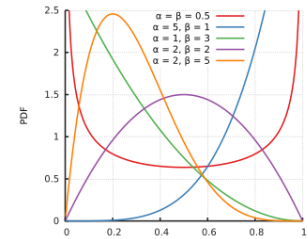
$$P(\theta_y | \mathbf{D}) \propto P(\mathbf{D} | \theta_y) P(\theta_y)$$

- Incorporating the Beta prior:

$$P(\theta) = \frac{\theta^{\alpha-1} (1-\theta)^{\beta-1}}{B(\alpha, \beta)}$$

$$\operatorname{argmax}_{\theta} P(\mathbf{D} | \theta_y) P(\theta_y) =$$

$$\operatorname{argmax}_{\theta} \log P(\mathbf{D} | \theta_y) + \log P(\theta_y)$$



20

## Intuition of the MAP result

$$p_y = \frac{|H| + (\alpha - 1)}{|H| + (\alpha - 1) + |T| + (\beta - 1)}$$

- Prior has strong influence when |H| and |T| small
- Prior has weak influence when |H| and |T| large

22

## Multinomial distribution

- What is mood of person in current minute?  $M = \{\text{Happy, Sad}\}$
- Measure his/her actions every ten seconds:  $A = \{\text{Cry, Jump, Laugh, Yell}\}$

Data ( $\mathbf{D}$ ): {LLJLCY, JLYJL, CCLLLJ, JJJJJ}Bias ( $\theta_y$ ): Probability table

	Happy	Sad
Cry	0.1	0.5
Jump	0.3	0.2
Laugh	0.5	0.1
Yell	0.1	0.3

$$P(\mathbf{D} | \theta_y) = (p_y^{\text{Cry}})^{|\text{Cry}|} (p_y^{\text{Jump}})^{|\text{Jump}|} (p_y^{\text{Laugh}})^{|\text{Laugh}|} (p_y^{\text{Yell}})^{|\text{Yell}|}$$

23

## Multinomial distribution – reexamined

$$P(\mathbf{D}|\boldsymbol{\theta}_y) = (p_y^{Cry})^{|Cry|} (p_y^{Jump})^{|Jump|} (p_y^{Laugh})^{|Laugh|} (p_y^{Yell})^{|Yell|}$$

More rigorously: in  $K$  measures,  $\delta(trial_k = \text{Action}) = \begin{cases} 0 & \text{if } trial_k \neq \text{Action} \\ 1 & \text{if } trial_k = \text{Action} \end{cases}$

$$P(\mathbf{D}|\boldsymbol{\theta}_y) = \prod_k \prod_i (p_y^{\text{Action}_i})^{\delta(trial_k = \text{Action}_i)}$$

24

## Learning parameters

$$\text{MLE: } P(A = a_i | M = m_j) = p_j^i = \frac{\#D\{A=a_i \wedge M=m_j\}}{\#D\{M=m_j\}}$$

$$\text{MAP: } P(A = a_i | M = m_j) = \frac{\#D(A=a_i \wedge M=m_j) + (\beta_j - 1)}{\#D(M=m_j) + \sum_m (\beta_m - 1)}$$

$$P(Y = y_j) = \frac{\#D(M=m_j) + (\beta_j - 1)}{|D| + \sum_m (\beta_m - 1)}$$

25

## Multiple multi-variate probabilities

Mood based on Action, Tunes, Weather

$$\text{argmax}_{\boldsymbol{\theta}_y} P(A, T, W | \boldsymbol{\theta}_y)$$

How many entries in probability table?

	Happy	Sad
Cry, Jazz, Sun	0.003	0.102
Cry, Jazz, Rain	0.024	0.025
	⋮	
Cry, Rap, Snow	0.011	0.115
	⋮	
Laugh, Rap, Rain	0.042	0.007
	⋮	
Yell, Opera, Wind	0.105	0.052

## Naïve bayes:

**Assuming independence of input features**

$$\text{argmax}_{\boldsymbol{\theta}_y} P(A, T, W | \boldsymbol{\theta}_y) =$$

$$\text{argmax}_{\boldsymbol{\theta}_y} P(A | \boldsymbol{\theta}_y) P(T | \boldsymbol{\theta}_y) P(W | \boldsymbol{\theta}_y)$$

How many entries in probability tables?

	Happy	Sad
Cry	0.1	0.5
Jump	0.3	0.2
Laugh	0.5	0.1
Yell	0.1	0.3

	Happy	Sad
Jazz	0.05	0.4
Rap	0.5	0.3
Opera	0.45	0.3

	Happy	Sad
Sun	0.6	0.2
Rain	0.05	0.3
Snow	0.3	0.3
Wind	0.05	0.2

28