Bayesian classification

CISC 5800

Professor Daniel Leeds

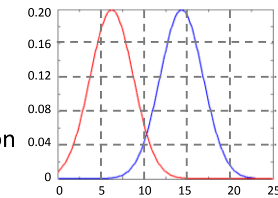
Classifying with probabilities

Example goal: Determine is it cloudy out

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

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

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

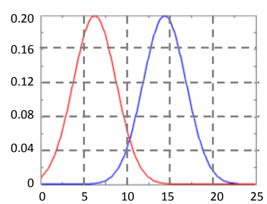


Classifying with probabilities

Example goal: Determine is it cloudy out

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

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



x=9

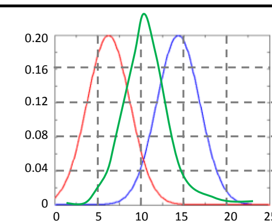
P(x=9|Cloudy)=0.12

P(x=9|Non-Cloud)=0.02

This is **Maximum Likelihood** classification

What if there's an eclipse?

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



x=9

P(x=9|Cloudy)=0.12

P(x=9|Non-Cloud)=0.02

P(x=9|Eclipse)=0.16

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 μ_y, σ_y
- **"Posterior probability"** estimated as likelihood × prior : $P\left(x|\mu_y, \sigma_y\right) P\left(\mu_y, \sigma_y\right)$
- Classify as $\text{argmax}_y P\left(x|\mu_y, \sigma_y\right) P\left(\mu_y, \sigma_y\right)$
- Terminology: μ_y, σ_y are "parameters." In general use $\boldsymbol{\theta_y}$
  Here: $\boldsymbol{\theta_y} = \{\mu_y, \sigma_y\}$. "**Posterior**" estimate is $P\left(x|\theta_y\right) P(\boldsymbol{\theta_y})$

7

Probability review: Bayes rule

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

and: $\mathrm{P}(A, B) = P(B|A)P(A)$

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

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

The **true** posterior

8

The posterior estimate

$$\underset{\boldsymbol{\theta_y}}{\text{argmax}}\, P(\boldsymbol{\theta_y}|\boldsymbol{D}) \propto P(\boldsymbol{D}|\boldsymbol{\theta_y})P(\boldsymbol{\theta_y})$$

Posterior ∝ Likelihood x Prior

∝ - means proportional

We "ignore" the P(**D**) denominator because **D** stays same while comparing different classes (*y* represented by $\boldsymbol{\theta_y}$)

9

Typical classification approaches

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

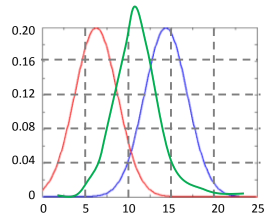
$$\underset{\boldsymbol{\theta_y}}{\text{argmax}}\, P(\boldsymbol{D}|\boldsymbol{\theta_y})$$

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

$$\underset{\boldsymbol{\theta_y}}{\text{argmax}}\, P(\boldsymbol{\theta_y}|\boldsymbol{D})$$

10

Incorporating a prior



Three classes:
Y={Cloudy, Non-Cloudy, Eclipse}

P(Cloudy)=0.4

P(Non-Cloudy)=0.4

P(Eclipse)=0.2

x=9

P(x=9|Cloudy) P(Cloud) =0.12x.4 = .048

P(x=9|Non-Cloud) P(Non-Cloud) = 0.02x.4 = 0.008

P(x=9|Eclipse) P(Eclipse)=0.16x.2 = .032

12

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 (**D**): {HHTH, TTHH, TTTT} Bias (θ_y): p_y probability of H for coin y

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

14

Bernoulli distribution – reexamined

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

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

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

15

Multinomial example

4-sided die – 4 probabilities:
$p_{side1}, p_{side2}, p_{side3}, p_{side4}$ (Note: $p_{side4} = 1 - \sum_{k=1}^{3} p_{sidek}$)

Define: $\delta(x) = \begin{cases} 1 & x = 0 \\ 0 & otherwise \end{cases}$

$$P(\boldsymbol{D}|\theta_y) = \prod_k p_{side1}^{\delta(side_k-1)} p_{side2}^{\delta(side_k-2)} p_{side3}^{\delta(side_k-3)} p_{side4}^{\delta(side_k-4)}$$

16

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

$\underset{\theta}{\text{argmax}}\, P(\boldsymbol{D}|\theta_y) =$

$\underset{p}{\text{argmax}}\, p_y^{|H|}(1-p_y)^{|T|}$

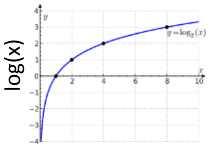
p_y - probability of Head

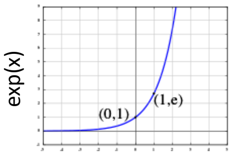
Equivalently, maximize $\log P(\boldsymbol{D}|\theta_y)$

$\underset{p_y}{\text{argmax}}\, |H|\log p_y + |T|\log(1-p_y)$

18

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

- 0.0000454 x 0.000912 = 0.0000000414 -> -10 + -7 = -17

19

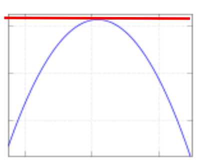
Optimization: finding zero slope

Location of maximum has slope 0

p - probability of Head

maximize $\log P(\boldsymbol{D}|\theta)$

$\underset{p}{\text{argmax}}\, |H|\log p + |T|\log(1-p):$



$\frac{d}{dp}|H|\log p + |T|\log(1-p) = 0$

$\frac{|H|}{p} - \frac{|T|}{1-p} = 0$

21

Intuition of the MLE result

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

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

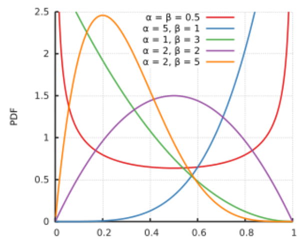
22

Finding the maximum **a posteriori**

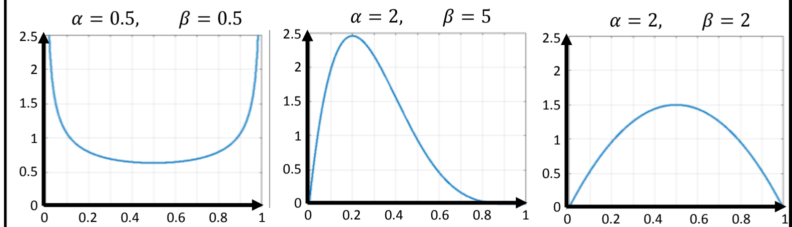
- $P(\theta_y|\boldsymbol{D}) \propto P(\boldsymbol{D}|\theta_y)P(\theta_y)$
- Incorporating the Beta prior:

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

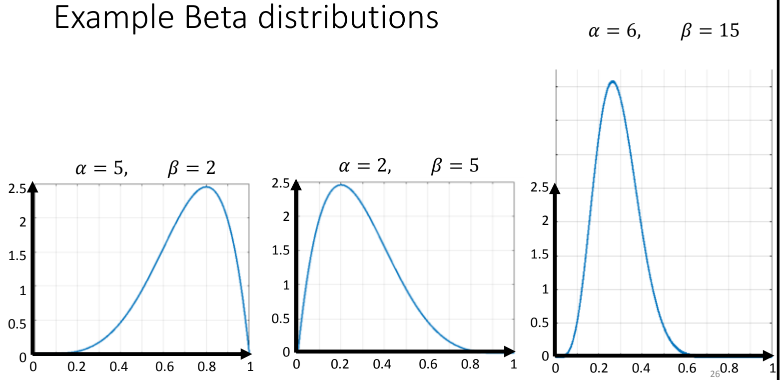
$$\underset{\theta}{\operatorname{argmax}} P(D|\theta_y)P(\theta_y) =$$
$$\underset{\theta}{\operatorname{argmax}} \log P(D|\theta_y) + \log P(\theta_y)$$



Example Beta distributions



Example Beta distributions



MAP: estimating θ (estimating p)

$$\underset{\theta}{\operatorname{argmax}} \log P(D|\theta) + \log P(\theta)$$
$$\underset{p}{\operatorname{argmax}} |H| \log p + |T| \log(1-p) +$$
$$(\alpha - 1)\log p + (\beta - 1)\log(1-p) - \log(\mathrm{B}(\alpha,\beta))$$

Set derivative to 0

$$\frac{|H|}{p} - \frac{|T|}{1-p} + \frac{(\alpha-1)}{p} - \frac{(\beta-1)}{1-p} = 0$$

$$(1-p)|H| - p|T| + (1-p)(\alpha-1) - p(\beta-1) = 0$$

$$|H| + (\alpha-1) = (|H| + |T| + (\alpha-1) + (\beta-1))p$$

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

- **$\alpha > \beta$ means expect to find coins biased to heads**
- **$\beta > \alpha$ means expect to find coins biased to tails**

28

Multinomial distribution Classification

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

Data (**D**): {LLJLCY, JJLYJL, CCLLLJ, JJJJJJ}

Bias ($\boldsymbol{\theta_y}$): Probability table

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

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

29

Multinomial distribution – reexamined

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

More rigorously: in K measures,

$$\delta(trial_k = \text{Action}) = \begin{cases} 0 & \text{if } trial_k \neq \text{Action} \\ 1 & \text{if } trial_k = \text{Action} \end{cases}$$

$$P(\boldsymbol{D}|\theta_y) = \prod_k \prod_i \left(p_y^{\text{Action}_i}\right)^{\delta(trial_k = \text{Action}_i)}$$

Classification: Given known likelihoods for each action, find mood that maximizes likelihood of observed sequence of actions (assuming each action is independent in the sequence)

30

Learning parameters

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

MAP: $P(A = a_i | M = m_j) = \frac{\#D(A=a_i \wedge M=m_j)+(\gamma_i - 1)}{\#D(M=m_j)+\sum_k(\gamma_k - 1)}$

γ_k is prior probability of each action class a_k

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

β_k is prior probability of each mood class m_k

32

Multiple multi-variate probabilities

Mood based on Action, Tunes, Weather

$$\underset{\boldsymbol{\theta_y}}{\operatorname{argmax}} P(A, T, W | \boldsymbol{\theta_y})$$

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

 # params = |M|x(|A|x|T|x|W|-1)

36

Naïve bayes:

Assuming independence of input features

$$\underset{\boldsymbol{\theta_y}}{\operatorname{argmax}} P(A, T, W | \boldsymbol{\theta_y}) =$$
$$\underset{\boldsymbol{\theta_y}}{\operatorname{argmax}} P(A | \boldsymbol{\theta_y}) P(T | \boldsymbol{\theta_y}) P(W | \boldsymbol{\theta_y})$$

How many entries in probability tables?

	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

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

- **# params = |M|x((|A|-1)+(|T|-1)+(|W|-1))** = 2x(3+2+3)=16

Benefits of Naïve Bayes

Very fast learning and classifying:

- **For multinomial problem:**
  - **Naïve independence: learn $|\mathbf{Y}| \times \sum_i (|\mathbf{X_i}| - 1)$ parameters**
  - **Non-naïve: learn $|\mathbf{Y}| \times (\prod_i |\mathbf{X_i}| - 1)$ parameters**

$|Y|$ is number of possible classes

$|X_i|$ is number of possible values for i[th] feature

Often works even if features are NOT independent

39

Typical Naïve Bayes classification

$$\underset{\boldsymbol{\theta_y}}{\operatorname{argmax}} P(\boldsymbol{\theta_y} | \boldsymbol{D}) \rightarrow \underset{\boldsymbol{\theta_y}}{\operatorname{argmax}} P(\boldsymbol{D} | \boldsymbol{\theta_y}) P(\boldsymbol{\theta_y})$$

$P(\boldsymbol{\theta_y})$ prior class probability

$$P(\boldsymbol{D} | \boldsymbol{\theta_y}) = \prod_i P(X^i | \boldsymbol{\theta_y})$$

where $\boldsymbol{D} = \begin{bmatrix} x^1 \\ \vdots \\ x^n \end{bmatrix}$ is a list of feature values

e.g., x^1=Action, x^2=Tunes

NB (Naïve Bayes): Find class y with $\boldsymbol{\theta_y}$ to maximize $P(\boldsymbol{\theta_y} | \boldsymbol{D})$

Assuming variables independent

40

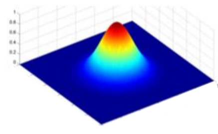
Multi-dimensional probability functions

- Multiple features as vector: $\mathbf{x} = \begin{bmatrix} temperature \\ windSpeed \\ musicVolume \end{bmatrix}$
- In 1D: likelihood P(temperature | mood)

$$L = \frac{\exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{x}-\boldsymbol{\mu})^2\right)}{\sigma\sqrt{2\pi}}$$

- In 2D: likelihood $P\left(\begin{bmatrix} temp \\ wind \end{bmatrix} | \, mood\right)$

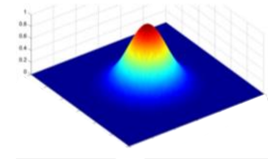
$$L = \frac{\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)}{\sqrt{(2\pi)^2|\boldsymbol{\Sigma}|}}$$



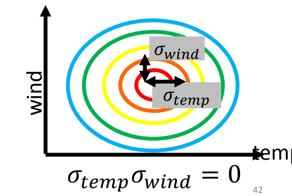
Multi-dimensional Gaussian – Naïve

- In 2D: likelihood $P\left(\begin{bmatrix} temp \\ wind \end{bmatrix} | \, mood\right)$

$$L = \frac{\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)}{\sqrt{(2\pi)^2|\boldsymbol{\Sigma}|}}$$



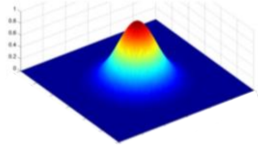
- $\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{temp}^2 & \sigma_{temp}\sigma_{wind} \\ \sigma_{temp}\sigma_{wind} & \sigma_{wind}^2 \end{bmatrix}$
- $\boldsymbol{\mu} = [\mu_{temp} \quad \mu_{wind}]$

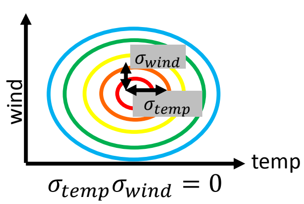


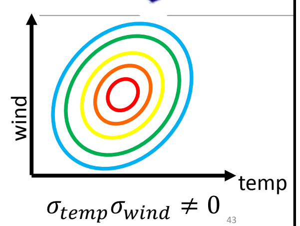
Multi-dimensional Gaussian – Non-naive

- In 2D: likelihood $P\left(\begin{bmatrix} temp \\ wind \end{bmatrix} | \, mood\right)$

$$\Sigma = \begin{bmatrix} \sigma_{temp}^2 & \sigma_{temp}\sigma_{wind} \\ \sigma_{temp}\sigma_{wind} & \sigma_{wind}^2 \end{bmatrix}$$







Gaussian parameter counts

For k dimensions

- Naïve: k + k $\approx 2k$ parameters
- Non-naïve: $k + \frac{k(k-1)}{2} \approx \frac{k^2}{2}$
- $\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \cdots & \sigma_1\sigma_k \\ \vdots & \ddots & \vdots \\ \sigma_1\sigma_k & \cdots & \sigma_k^2 \end{bmatrix}$