

Inf2b - Learning

Lecture 6: Naive Bayes

Hiroshi Shimodaira

(Credit: Iain Murray and Steve Renals)

Centre for Speech Technology Research (CSTR)

School of Informatics

University of Edinburgh

<http://www.inf.ed.ac.uk/teaching/courses/inf2b/>

<https://piazza.com/ed.ac.uk/spring2020/infr08028>

Office hours: Wednesdays at 14:00-15:00 in IF-3.04

Jan-Mar 2020

Today's Schedule

- 1 Bayes decision rule review
- 2 The curse of dimensionality
- 3 Naive Bayes
- 4 Text classification using Naive Bayes (introduction)

Bayes decision rule (recap)

Class $C = \{1, \dots, K\}$; C_k to denote $C = k$; input features $X = \mathbf{x}$

Most probable class: (maximum posterior class)

$$\begin{aligned}k_{\max} &= \arg \max_{k \in C} P(C_k | \mathbf{x}) = \arg \max_k \frac{P(\mathbf{x} | C_k) P(C_k)}{\sum_{j=1}^K P(\mathbf{x} | C_j) P(C_j)} \\ &= \arg \max_k P(\mathbf{x} | C_k) P(C_k)\end{aligned}$$

where $P(C_k | \mathbf{x})$: posterior
 $P(\mathbf{x} | C_k)$: likelihood
 $P(C_k)$: prior

⇒ Minimum error (misclassification) rate classification

(PRML C. M. Bishop (2006) Section 1.5)

Fish classification (revisited)

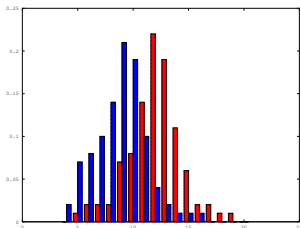
Bayesian class estimation:

$$P(C_k | x) = \frac{P(x | C_k) P(C_k)}{P(x)} \propto P(x | C_k) P(C_k)$$

Estimating the terms: (Non-Bayesian)

$$\text{Priors: } P(C = M) \approx \frac{N_M}{N_M + N_F}, \dots$$

$$\text{Likelihoods: } P(x | C = M) \approx \frac{n_M(x)}{N_M}, \dots$$



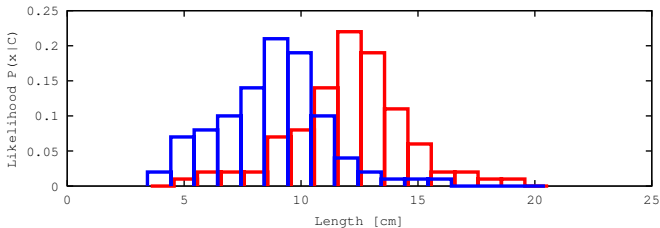
NB: These *approximations* work well only if we have enough data

Fish classification (revisited)

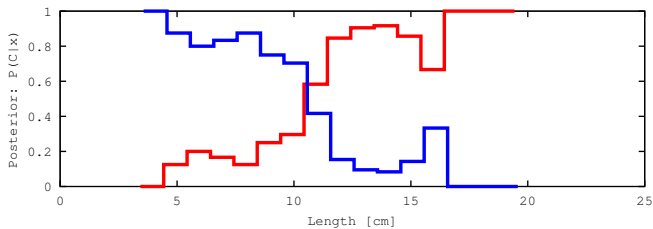
$$P(C_k | x) = \frac{P(x | C_k) P(C_k)}{P(x)}$$

$$P(M) : P(F) = 1 : 1$$

$P(x | C_k)$



$P(C_k | x)$

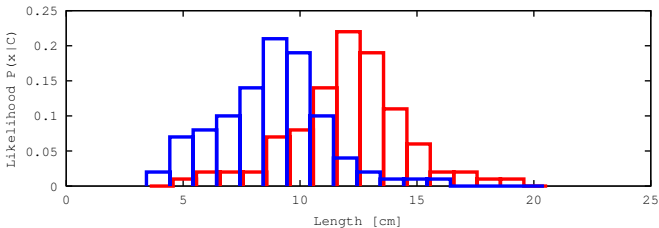


Fish classification (revisited)

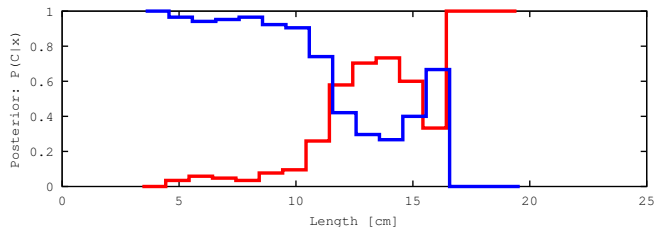
$$P(C_k | x) = \frac{P(x | C_k) P(C_k)}{P(x)}$$

$$P(M) : P(F) = 1 : 4$$

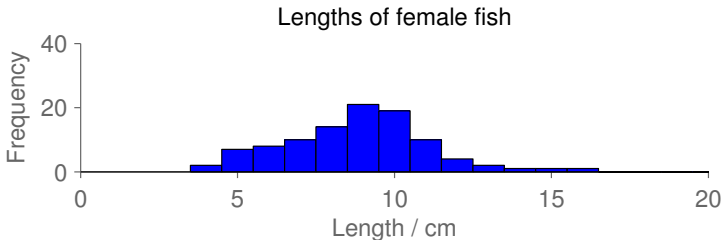
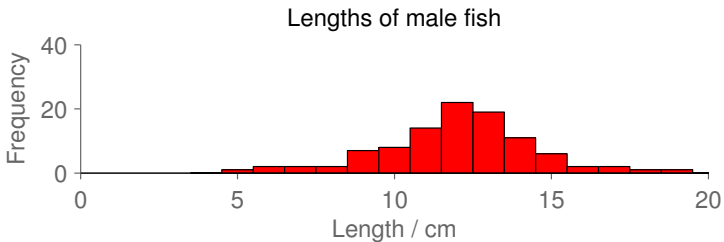
$P(x | C_k)$



$P(C_k | x)$



How can we improve the fish classification?



More features!?

$$P(\mathbf{x} | C_k) \approx \frac{n_{C_k}(x_1, \dots, x_D)}{N_{C_k}}$$

1D histogram: $n_{C_k}(x_1)$

2D histogram: $n_{C_k}(x_1, x_2)$

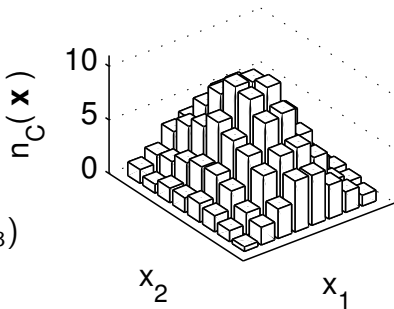
3D cube of numbers: $n_{C_k}(x_1, x_2, x_3)$

⋮

100 binary variables, 2^{100} settings (the universe is $\approx 2^{98}$ picoseconds old)

In high dimensions almost all $n_{C_k}(x_1, \dots, x_D)$ are zero

⇒ Bellman's "curse of dimensionality"



Avoiding the Curse of Dimensionality

Apply the chain rule?

$$\begin{aligned}P(\mathbf{x} | C_k) &= P(x_1, x_2, \dots, x_D | C_k) \\ &= P(x_1 | C_k) P(x_2 | x_1, C_k) P(x_3 | x_2, x_1, C_k) P(x_4 | x_3, x_2, x_1, C_k) \cdots \\ &\quad \cdots P(x_{d-1} | x_{d-2}, \dots, x_1, C_k) P(x_D | x_{D-1}, \dots, x_1, C_k)\end{aligned}$$

Solution: assume structure in $P(\mathbf{x} | C_k)$

For example,

- Assume x_{d+1} depends on x_d only

$$P(\mathbf{x} | C_k) \approx P(x_1 | C_k) P(x_2 | x_1, C_k) P(x_3 | x_2, C_k) \cdots P(x_D | x_{D-1}, C_k)$$

- Assume $\mathbf{x} \in \mathcal{R}^D$ distributes in a low dimensional vector space
 - Dimensionality reduction by PCA (Principal Component Analysis) / KL-transform

Avoiding the Curse of Dimensionality (*cont.*)

- Apply smoothing windows (e.g. Parzen windows)
- Apply a probability distribution model (e.g. Normal dist.)
- Assume x_1, \dots, x_D are **conditionally independent** given class

⇒ **Naive Bayes** rule/model/assumption
(or *idiot Bayes rule*)

$$\begin{aligned} P(x_1, x_2, \dots, x_D | C_k) &= P(x_1 | C_k) P(x_2 | C_k) \cdots P(x_D | C_k) \\ &= \prod_{d=1}^D P(x_d | C_k) \end{aligned}$$

- *Is it reasonable?*
Often not, of course!
Although it can still be *useful*.

Example - game played depending on the weather

| Outlook | Temperature | Humidity | Windy | Play |
|----------|-------------|----------|-------|------|
| sunny | hot | high | false | NO |
| sunny | hot | high | true | NO |
| overcast | hot | high | false | YES |
| rainy | mild | high | false | YES |
| rainy | cool | normal | false | YES |
| rainy | cool | normal | true | NO |
| overcast | cool | normal | true | YES |
| sunny | mild | high | false | NO |
| sunny | cool | normal | false | YES |
| rainy | mild | normal | false | YES |
| sunny | mild | normal | true | YES |
| overcast | mild | high | true | YES |
| overcast | hot | normal | false | YES |
| rainy | mild | high | true | NO |

$$P(\text{Play} \mid O, T, H, W) = \frac{P(O, T, H, W \mid \text{Play}) P(\text{Play})}{P(O, T, H, W)}$$

Weather data - how to calculate probabilities?

$$P(\text{Play} \mid O, T, H, W) = \frac{P(O, T, H, W \mid \text{Play}) P(\text{Play})}{P(O, T, H, W)}$$

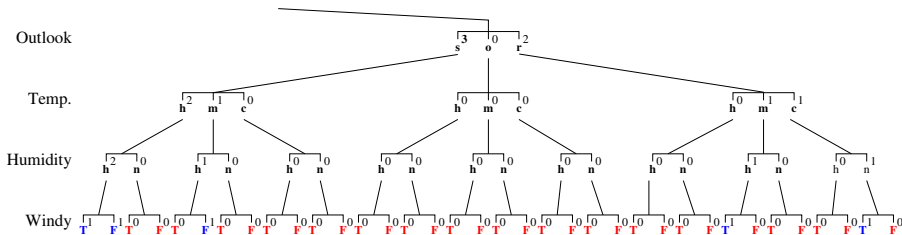
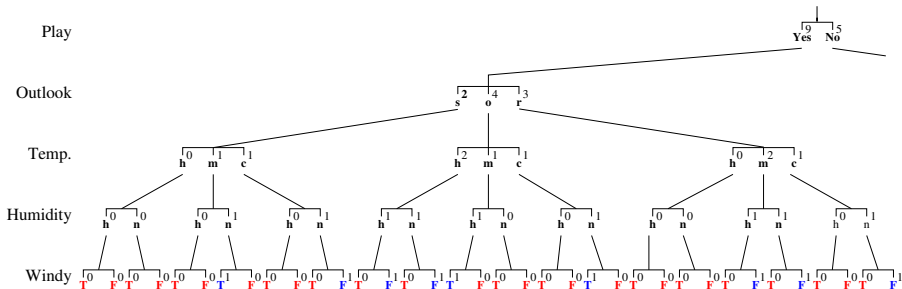
If we use histograms for this
4D data: $n_{\text{Play}}(O, T, H, W)$

$$\begin{array}{c} \text{Outlook} \\ \left[\begin{array}{c} \text{sunny} \\ \text{overcast} \\ \text{rainy} \end{array} \right] \end{array} \times \begin{array}{c} \text{Temp.} \\ \left[\begin{array}{c} \text{hot} \\ \text{mild} \\ \text{cool} \end{array} \right] \end{array} \times \begin{array}{c} \text{Humidity} \\ \left[\begin{array}{c} \text{high} \\ \text{normal} \end{array} \right] \end{array} \times \begin{array}{c} \text{Windy} \\ \left[\begin{array}{c} \text{true} \\ \text{false} \end{array} \right] \end{array}$$

of bins in the histogram = $3 \times 3 \times 2 \times 2 = 36$

of samples available = 9 for play:yes, 5 for play:no

Weather data - tree representation



Applying Naive Bayes

$$P(\text{Play} \mid O, T, H, W) = \frac{P(O, T, H, W \mid \text{Play}) P(\text{Play})}{P(O, T, H, W)}$$
$$\propto P(O, T, H, W \mid \text{Play}) P(\text{Play})$$

Applying the Naive Bayes rule,

$$P(O, T, H, W \mid \text{Play}) = P(O \mid \text{Play}) P(T \mid \text{Play}) P(H \mid \text{Play}) P(W \mid \text{Play})$$

Weather data summary

Counts

| | Outlook | | Temperature | | Humidity | | Windy | | Play | | | | |
|-------|---------|---|-------------|---|----------|------|-------|---|------|---|---|---|---|
| | Y | N | Y | N | Y | N | Y | N | Y | N | | | |
| sunny | 2 | 3 | hot | 2 | 2 | high | 3 | 4 | f | 6 | 2 | 9 | 5 |
| overc | 4 | 0 | mild | 4 | 2 | norm | 6 | 1 | t | 3 | 3 | | |
| rainy | 3 | 2 | cool | 3 | 1 | | | | | | | | |

Relative frequencies $P(x|Play = Y), P(x|Play = N)$

| | Outlook | | Temperature | | Humidity | | Windy | | Play | | | | |
|---|---------|-----|-------------|-----|----------|---|-------|-----|------|------|-----|------|------|
| | Y | N | Y | N | Y | N | Y | N | P(Y) | P(N) | | | |
| s | 2/9 | 3/5 | h | 2/9 | 2/5 | h | 3/9 | 4/5 | f | 6/9 | 2/5 | 9/14 | 5/14 |
| o | 4/9 | 0/5 | m | 4/9 | 2/5 | n | 6/9 | 1/5 | t | 3/9 | 3/5 | | |
| r | 3/9 | 2/5 | c | 3/9 | 1/5 | | | | | | | | |

Test example

$x =$ (Outlook Temp. Humidity Windy Play
 (sunny cool high true) ?)

Weather data summary (Ver.2)

Counts

| Play | Outlook | | | Temp. | | | Humidity | | Windy | |
|-------|---------|-------|-------|-------|------|------|----------|------|-------|------|
| | sunny | overc | rainy | hot | mild | cool | high | norm | False | True |
| Yes 9 | 2 | 4 | 3 | 2 | 4 | 3 | 3 | 6 | 6 | 3 |
| No 5 | 3 | 0 | 2 | 2 | 2 | 1 | 4 | 1 | 2 | 3 |

Relative frequencies $P(x|Play)$

| Play | $P(Play)$ | Outlook | | | Temp. | | | Humidity | | Windy | |
|------|-----------|---------|-------|-------|-------|------|------|----------|------|-------|------|
| | | sunny | overc | rainy | hot | mild | cool | high | norm | False | True |
| Y | 9/14 | 2/9 | 4/9 | 3/9 | 2/9 | 4/9 | 3/9 | 3/9 | 6/9 | 6/9 | 3/9 |
| N | 5/14 | 3/5 | 0/5 | 2/5 | 2/5 | 2/5 | 1/5 | 4/5 | 1/5 | 2/5 | 3/5 |

Test example

$\mathbf{x} =$ Outlook Temp. Humidity Windy Play
 (sunny cool high true) ?

Applying Naive Bayes

Posterior prob. of "play" given $\mathbf{x} = (\text{sunny, cool, humid, windy})$

$$P(\text{Play} | \mathbf{x}) \propto P(\mathbf{x} | \text{Play}) P(\text{Play})$$

$$\begin{aligned} P(\text{Play} = Y | \mathbf{x}) &\propto P(O=s|Y) P(T=c|Y) P(H=h|Y) P(W=t|Y) P(Y) \\ &\propto \frac{2}{9} \cdot \frac{3}{9} \cdot \frac{3}{9} \cdot \frac{3}{9} \cdot \frac{9}{14} \approx 0.0053 \end{aligned}$$

$$\begin{aligned} P(\text{Play} = N | \mathbf{x}) &\propto P(O=s|N) P(T=c|N) P(H=h|N) P(W=t|N) P(N) \\ &\propto \frac{3}{5} \cdot \frac{1}{5} \cdot \frac{4}{5} \cdot \frac{3}{5} \cdot \frac{5}{14} \approx 0.0206 \end{aligned}$$

Exercise: find the odds of play, $P(\text{play} = Y | \mathbf{x}) / P(\text{play} = N | \mathbf{x})$
(answer in notes)

Easy and cheap:

Record counts, convert to frequencies, score each class by multiplying prior and likelihood terms

$$P(C_k | \mathbf{x}) \propto \left(\prod_{d=1}^D P(x_d | C_k) \right) P(C_k)$$

Statistically viable:

Simple count-based estimates work in 1D

Often overconfident:

Treats dependent evidence as independent

Another approach for the weather example

- What about applying k -NN?
- Data representation (by quantification)

$$X = \begin{pmatrix} O & T & H & W & P \\ 3 & 3 & 2 & 0 & 0 \\ 3 & 3 & 2 & 1 & 0 \\ 2 & 3 & 2 & 0 & 1 \\ 1 & 2 & 2 & 0 & 1 \\ 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 \\ 2 & 1 & 1 & 1 & 1 \\ 3 & 2 & 2 & 0 & 0 \\ 3 & 1 & 1 & 0 & 1 \\ 1 & 2 & 1 & 0 & 1 \\ 3 & 2 & 1 & 1 & 1 \\ 2 & 2 & 2 & 1 & 1 \\ 2 & 3 & 1 & 0 & 1 \\ 1 & 2 & 2 & 1 & 0 \end{pmatrix}$$

$$\mathbf{x} = (3 \quad 1 \quad 2 \quad 1)$$

| | | |
|---------|-------|---|
| Outlook | sunny | 3 |
| | overc | 2 |
| | rainy | 1 |
| Temp. | hot | 3 |
| | mild | 2 |
| | cold | 1 |
| Humid. | high | 2 |
| | norm | 1 |
| Windy | True | 1 |
| | False | 0 |
| Play | Yes | 1 |
| | No | 0 |

Another approach for the weather example (*cont.*)

- k -NN
- Sorted distance between $X(:, 1:4)$ and \mathbf{x}

| rank | dist. | idx | label | rank | dist. | idx | label |
|------|-------|------|-------|------|-------|------|-------|
| 1 | 1.41 | (7) | Y | 1 | 1.41 | (8) | N |
| 2 | 1.41 | (8) | N | 2 | 1.41 | (12) | Y |
| 3 | 1.41 | (9) | Y | 3 | 2.00 | (2) | N |
| 4 | 1.41 | (11) | Y | 4 | 2.24 | (1) | N |
| 5 | 1.41 | (12) | Y | 5 | 2.24 | (7) | Y |
| 6 | 2.00 | (2) | N | 6 | 2.24 | (9) | Y |
| 7 | 2.24 | (1) | N | 7 | 2.24 | (11) | Y |
| 8 | 2.24 | (6) | N | 8 | 2.24 | (14) | N |
| 9 | 2.24 | (14) | N | 9 | 2.45 | (3) | Y |
| 10 | 2.45 | (3) | Y | 10 | 2.45 | (4) | Y |
| 11 | 2.45 | (4) | Y | 11 | 2.83 | (6) | N |
| 12 | 2.45 | (5) | Y | 12 | 3.00 | (5) | Y |
| 13 | 2.65 | (10) | Y | 13 | 3.16 | (10) | Y |
| 14 | 2.65 | (13) | Y | 14 | 3.16 | (13) | Y |

where the values for Humidity were doubled.

Another approach for the weather example (*cont.*)

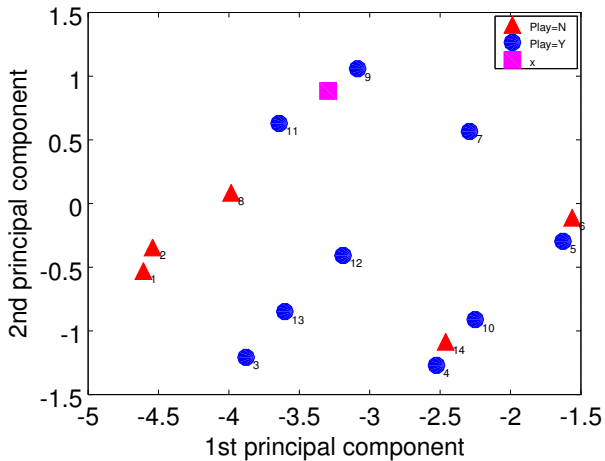
- Correlation matrix for (O, T, H, W, P)

| | O | T | H | W | P |
|-----|----------|----------|----------|----------|----------|
| O | 1.00000 | 0.33541 | 0.16903 | 0.00000 | -0.17638 |
| T | 0.33541 | 1.00000 | 0.56695 | -0.19094 | -0.19720 |
| H | 0.16903 | 0.56695 | 1.00000 | 0.00000 | -0.44721 |
| W | 0.00000 | -0.19094 | 0.00000 | 1.00000 | -0.25820 |
| P | -0.17638 | -0.19720 | -0.44721 | -0.25820 | 1.00000 |

NB: Humidity has the largest (negative) correlation with Play.

Another approach for the weather example (*cont.*)

- Dimensionality reduction by PCA



Exercise (past exam question)

The table gives a small dataset. Tick marks indicate which movies 3 children (marked c) and 4 adults (marked a) have watched. The final two rows give the movies watched by two users of the system of unknown age.

| type | m_1 | m_2 | m_3 | m_4 |
|-------|-------|-------|-------|-------|
| c | ✓ | | | |
| c | ✓ | | ✓ | |
| c | ✓ | | | |
| a | | ✓ | | |
| a | | | ✓ | |
| a | ✓ | ✓ | ✓ | ✓ |
| y_1 | ✓ | | ✓ | |
| y_2 | | | ✓ | |

Apply maximum likelihood estimation of the priors and likelihoods to this data, using the naive Bayes assumption for the likelihoods. Hence find the odds that the test user y_i is child: $P(y_i = c|\text{data})/P(y_i = a|\text{data})$ for $i = 1, 2$. State the MAP classification of each user.

Spam?

I got your contact information from your country's information directory during my desperate search for someone who can assist me secretly and confidentially in relocating and managing some family fortunes.

Spam?

Dear Dr. Steve Renals, The proof for your article, Combining Spectral Representations for Large-Vocabulary Continuous Speech Recognition, is ready for your review. Please access your proof via the user ID and password provided below. Kindly log in to the website within 48 HOURS of receiving this message so that we may expedite the publication process.

Identifying Spam

Spam?

Congratulations to you as we bring to your notice, the results of the First Category draws of THE HOLLAND CASINO LOTTO PROMO INT. We are happy to inform you that you have emerged a winner under the First Category, which is part of our promotional draws.

Identifying Spam

Question

How can we identify an email as spam automatically?

Text classification: classify email messages as spam or non-spam (ham), based on the words they contain

With the Bayes decision rule,

$$P(\text{Spam}|\mathbf{x}_1, \dots, \mathbf{x}_L) \propto P(\mathbf{x}_1, \dots, \mathbf{x}_L|\text{Spam})P(\text{Spam})$$

Using the naive Bayes assumption,

$$P(\mathbf{x}_1, \dots, \mathbf{x}_L|\text{Spam}) = P(\mathbf{x}_1|\text{Spam}) \cdots P(\mathbf{x}_L|\text{Spam})$$

Summary

- The curse of dimensionality
- Approximation by the Naive Bayes rule
- Example: classifying multidimensional data using Naive Bayes
- Next lecture: Text classification using Naive Bayes