Inf2b Learning and Data

Lecture 6: Naive Bayes

Hiroshi Shimodaira
(Credit: Iain Murray and Steve Renals)

Centre for Speech Technology Research (CSTR)
School of Informatics
University of Edinburgh

Jan-Mar 2014

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 $C = \{c_1, \ldots, c_K\}$; input features $X = \mathbf{x}$

Most probable class: (maximum posterior class)

$$
\begin{aligned}
c^* &= \arg\max_{c_k} \; P(c_k \,|\, \mathbf{x}) \\
&= \arg\max_{c_k} \frac{P(\mathbf{x} \,|\, c_k) \; P(c_k)}{P(\mathbf{x})} \;=\; \arg\max_{c_k} \frac{P(\mathbf{x} \,|\, c_k) \; P(c_k)}{\sum_{j=1}^{K} P(\mathbf{x}|c_j) P(c_j)} \\
&= \arg\max_{c_k} \; P(\mathbf{x}|c_k) P(c_k)
\end{aligned}
$$

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

$\Rightarrow$ Minimum error (misclassification) rate classification

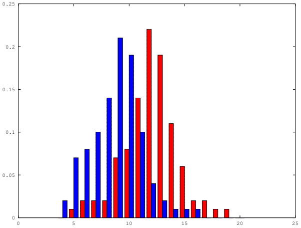
(PRML C. M. Bishop (2006) Section 1.5)

Fish classification (revisited)

Bayesian class estimation:

$$P(c_k \mid x) = \frac{P(x \mid c_k)\, P(c_k)}{P(x)} \propto P(x \mid c_k)\, P(c_k)$$

Estimating the terms: (Non-Bayesian)



Priors: $P(C = M) \approx \dfrac{N_M}{N_M + N_F}, \ldots$

Likelihoods: $P(x \mid C = M) \approx \dfrac{n_M(x)}{N_M}, \ldots$

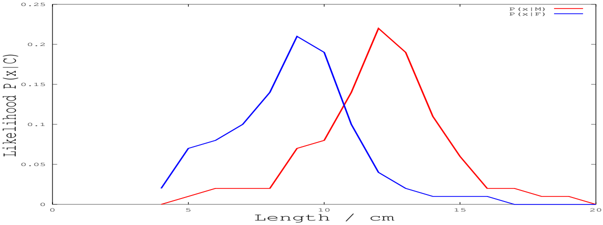
NB: These *approximations* work well only if we have enough data

Fish classification (revisited)

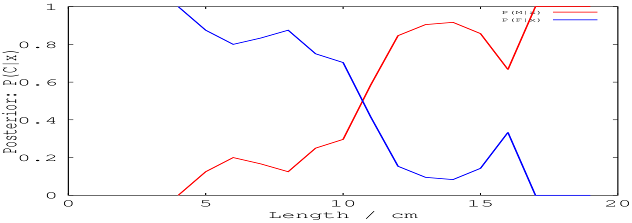
$$P(c_k|x) = \frac{P(x|c_k)P(c_k)}{P(x)}$$

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

$P(x|c_k)$



$P(c_k|x)$

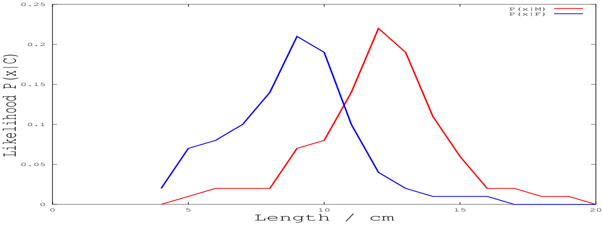


Fish classification (revisited)

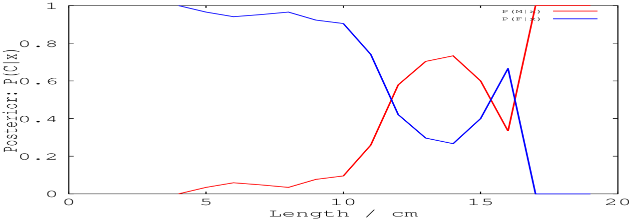
$$P(c_k|x) = \frac{P(x|c_k)P(c_k)}{P(x)}$$

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

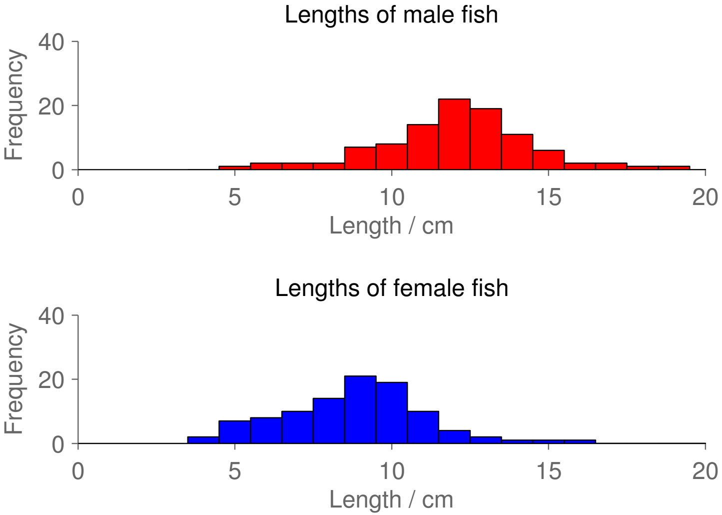
$P(x|c_k)$



$P(c_k|x)$



How can we improve the fish classification?



More features!?

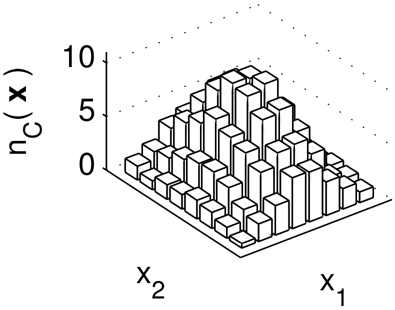
$$P(\mathbf{x} \mid c_k) \approx \frac{n_{c_k}(x_1, \ldots, x_d)}{N_{c_k}}$$

1D histogram

2D histogram

3D cube of numbers

$\vdots$



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

In high dimensions almost all $n_C(x_1, \ldots, x_D)$ are zero

$\Rightarrow$ Bellman's "curse of dimensionality"

Avoiding the Curse of Dimensionality

Apply the chain rule?

$$
\begin{aligned}
P(\mathbf{x} \,|\, c_k) &= P(x_1, x_2, \ldots, x_d \,|\, c_k) \\
&= P(x_1|c_k)\, P(x_2|x_1, c_k)\, P(x_3|x_2, x_1, c_k)\, P(x_4|x_3, x_2, x_1, c_k) \cdots \\
&\qquad \cdots P(x_{d-1}|x_{d-2}, \ldots, x_1, c_k)\, P(x_d|x_{d-1}, \ldots, x_1, c_k)
\end{aligned}
$$

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

For example,

- Assume x_{i+1} depends on x_i 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

- Apply smoothing windows (e.g. Parzen windows)
- Apply a probability distribution model (e.g. Normal dist.)
- Assume $x_1, \ldots, x_d$ are independent from each other

  $\Rightarrow$ Naive Bayes rule/model (or *idiot Bayes rule)*

$$P(x_1, x_2, \ldots, x_d | c_k) \approx P(x_1|c_k)\, P(x_2|c_k) \cdots P(x_d|c_k) = \prod_{i=1}^{d} P(x_i|c_k)$$

- *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

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

# of combinations of $(O,T,H,W) = 3 \times 3 \times 2 \times 2 = 36$

Applying Naive Bayes

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

$$\propto P(O,T,H,W|Play)\,P(Play)$$

Applying the Naive Bayes rule,

$$P(O,T,H,W|Play) \approx P(O|Play)\,P(T|Play)\,P(H|Play)\,P(W|Play)$$

Relative frequencies

Consider each feature independently
to estimate $P(O|Play), P(T|Play), P(H|Play), P(W|Play)$

Outlook	**Y**	**N**
sunny	2/9	3/5
overcast	4/9	0/5
rainy	3/9	2/5

Temperature	**Y**	**N**
hot	2/9	2/5
mild	4/9	2/5
cool	3/9	1/5

Humidity	**Y**	**N**
high	3/9	4/5
normal	6/9	1/5

Windy	**Y**	**N**
false	6/9	2/5
true	3/9	3/5

There was play 9 out of 14 times: $P(\text{Play}=Y) \approx \frac{9}{14}$

Applying Naive Bayes

Posterior play probability: $\mathbf{x} = (\text{sunny, cool, humid, windy})$

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

Estimating the Naive Bayes likelihood: (Non-Bayesian)

$$P(\mathbf{x} \,|\, \text{Play}=Y) = P(O=s \,|\, Y)\, P(T=c \,|\, Y)\, P(H=h \,|\, Y)\, P(W=t \,|\, Y)$$
$$\approx \frac{2}{9} \cdot \frac{3}{9} \cdot \frac{3}{9} \cdot \frac{3}{9}$$

$$P(\mathbf{x} \,|\, \text{Play}=N) = P(O=s \,|\, N)\, P(T=c \,|\, N)\, P(H=h \,|\, N)\, P(W=t \,|\, N)$$
$$\approx \frac{3}{5} \cdot \frac{1}{5} \cdot \frac{4}{5} \cdot \frac{3}{5}$$

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

(answer in notes)

Naive Bayes properties

Easy and cheap:

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

$$P(\mathbf{x}|c_k) \propto \left(\prod_{i=1}^{d} P(x_i|c_k)\right) P(c_k)$$

Statistically viable:

Simple count-based estimates work in 1D

Often overconfident:

Treats dependent evidence as independent

Identifying Spam

Spam?

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

Identifying Spam

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

Text Classification using Bayes Theorem

- Document D, with class c_k
- Classify D as the class with the highest posterior probability:

$$P(c_k|D) = \frac{P(D|c_k)P(c_k)}{P(D)} \propto P(D|c_k)P(c_k)$$

- How do we represent D? How do we estimate $P(D|c_k)$?
- **Bernoulli document model**: a document is represented by a binary feature vector, whose elements indicate absence or presence of corresponding word in the document
- **Multinomial document model**: a document is represented by an integer feature vector, whose elements indicate frequency of corresponding word in the document

Summary

- The curse of dimensionality
- Naive Bayes approximation
- Example: classifying multidimensional data using Naive Bayes
- Next lecture: Text classification using Naive Bayes