COMPSCI762: Foundations of Machine Learning

Introduction to Bayesian Learning

Jörg Simon Wicker and Katerina Taškova

The University of Auckland

SCIENCE
SCHOOL OF COMPUTER SCIENCE

Bayesian Learning

Bayes Optimal Classifier
Naive Bayes Classifier
Naive Bayes for Document Classification
Bayesian Networks

Partly based on Mitchel's book, lecture slides from Stanford's NLP lecture and The University of Utah

Bayes Optimal Classifier

Bayes Optimal Classifier

- Question: What is the most probable classification of a new instance given the training data?
- Simply applying h_{MAP} is not the best solution (as one could wrongly think of)
- Example
  - $H = \{h_1, h_2, h_3\}$, where $P(h_1|D) = 0.4$, $P(h_2|D) = P(h_3|D) = 0.3$
  - $h_{MAP} = h_1$
  - Consider a new instance x encountered, which is classified positive by h_1 and negative by h_2 , h_3
  - Taking all hypotheses into account
    - The probability that x is positive is 0.4
    - The probability that x is negative is 0.6

$\Rightarrow$ most probable classification $\neq$ classification generated by h_{MAP}

Bayes Optimal Classifier

- The most probable classification is obtained by combining the predictions of all hypotheses, weighted by their posterior probabilities

$$P(v_j|D) = \sum_{h_i \in H} P(v_j|h_i)P(h_i|D)$$

where $P(v_j|D)$ is the probability that the correct classification is v_j

- Bayes optimal classifier

$$\arg\max_{v_j \in V} P(v_j|D) = \arg\max_{v_j \in V} \sum_{h_i \in H} P(v_j|h_i)P(h_i|D)$$

Example

$V = \{\oplus, \ominus\}$

$$P(h_1|D) = 0.4, P(\ominus|h_1) = 0, P(\oplus|h_1) = 1$$
$$P(h_2|D) = 0.3, P(\ominus|h_2) = 1, P(\oplus|h_2) = 0$$
$$P(h_3|D) = 0.3, P(\ominus|h_3) = 1, P(\oplus|h_3) = 0$$

therefore

$$P(\oplus|D) = \sum_{h_i \in H} P(\oplus|h_i)P(h_i|D) = 0.4$$
$$P(\ominus|D) = \sum_{h_i \in H} P(\ominus|h_i)P(h_i|D) = 0.6$$

and

$$\underset{v_j \in \{\oplus, \ominus\}}{\arg\max} \sum_{h_i \in H} P(v_j|h_i)P(h_i|D) = \ominus$$

Naive Bayes Classifier

Naive Bayes Classifier

- Applies to learning tasks where each instance x is described by a conjunction of attribute values $< a_1, a_2, \ldots, a_n >$ and where the target function $f(x)$ can take on any value from some finite set V
- The most probable target value is

$$\begin{aligned} v_{MAP} &= \arg\max_{v_j \in V} P(v_j | a_1, a_2, \ldots, a_n) \\ &= \arg\max_{v_j \in V} \frac{P(a_1, a_2, \ldots, a_n | v_j) P(v_j)}{P(a_1, a_2, \ldots, a_n)} \\ &= \arg\max_{v_j \in V} P(a_1, a_2, \ldots, a_n | v_j) P(v_j) \end{aligned}$$

Naive Bayes Classifier

- Given training data D, $P(v_j)$ can be estimated by counting the frequency of v_j in D
- However, it is not feasible to estimate $P(a_1, a_2, \ldots, a_n | v_j)$
  - Number of these terms is $|$all possible instances$| \times |V|$
- Simplification: naive Bayes classifier
  - Assumption: Attribute values are conditionally independent given the target value
  - **Absolute independence of X and Y**
    $P(X, Y) = P(X|Y)P(Y) = P(X)P(Y)$
  - **Conditional independence of X and Y given Z**
    $P(X, Y|Z) = P(X|Y, Z)P(Y|Z) = P(X|Z)P(Y|Z)$

Naive Bayes Classifier

- Given training data D, $P(v_j)$ can be estimated by counting the frequency of v_j in D
- However, it is not feasible to estimate $P(a_1, a_2, \ldots, a_n | v_j)$
  - Number of these terms is $|\text{all possible instances}| \times |V|$
- Simplification: naive Bayes classifier
  - Assumption:Attribute values are conditionally independent given the target value
  - Hence, $P(a_1, a_2, \ldots, a_n | v_j) = \prod_i P(a_i | v_j)$
  - Hence, number of terms is $|\text{distinct attribute values}| \times |V| + |V|$
  - No explicit search through hypothesis space H, just counting frequencies

⇒ If the above assumption is correct, Naive Bayes classifications are MAP classifications

$$v_{NB} = \underset{v_j \in V}{\arg\max}\, P(v_j) \prod_i P(a_i | v_j)$$

Naive Bayes Classifier - Example

We want to learn to discriminate object labeled + from objects labeled - using measurements for three categorical features. You are given the following objects.

ID	Heavy H	Size S	Colour C	Label L
Train1	Yes	L	R	+
Train2	No	M	B	+
Train3	No	L	G	+
Train4	No	L	G	+
Train5	Yes	M	R	-
Train6	Yes	L	G	-
Train7	No	M	G	-
Test1	Yes	M	R	
Test2	Yes	M	B	
Test3	Yes	M	?	

- Assume the features describing the object are independent of its class
- Calculate two posterior probabilities for each test example, before we can make decision
- Test1 example

$$
\begin{aligned}
&\frac{P(L=+|H=Yes,S=M,C=R)}{P(L=-|H=Yes,S=M,C=R)}\\
&=\frac{P(H=Yes,S=M,C=R|L=+)\cdot P(L=+)}{P(H=Yes,S=M,C=R|L=-)\cdot P(L=-)}\\
&=\frac{P(H=Yes|L=+)P(S=M|L=+)P(C=R|L=+)\cdot P(L=+)}{P(H=Yes|L=-)P(S=M|L=-)P(C=R|L=-)\cdot P(L=-)}
\end{aligned}
$$

Naive Bayes Classifier - Example (cont.)

Training phase outputs following probability estimates (tables bellow) used in **testing phase** to classify new examples.

$P(H\|L)$	L = +	L = -
H = Yes	$\frac{1}{4}$	$\frac{2}{3}$
H = No	$\frac{3}{4}$	$\frac{1}{3}$

$P(S\|L)$	L = +	L = -
S = L	$\frac{3}{4}$	$\frac{1}{3}$
S = M	$\frac{1}{4}$	$\frac{2}{3}$

$P(C\|L)$	L = +	L = -
C = R	$\frac{1}{4}$	$\frac{1}{3}$
C = B	$\frac{1}{4}$	0
C = G	$\frac{2}{4}$	$\frac{2}{3}$

	L = +	L = -
$P(L)$	$\frac{4}{7}$	$\frac{3}{7}$

- Test1 example

$$\frac{P(L=+|H=Yes,S=M,C=R)}{P(L=-|H=Yes,S=M,C=R)}$$
$$=\frac{P(H=Yes|L=+)P(S=M|L=+)P(C=R|L=+)\cdot P(L=+)}{P(H=Yes|L=-)P(S=M|L=-)P(C=R|L=-)\cdot P(L=-)}$$
$$=\frac{9}{64}<1$$

$\Rightarrow$ Test1 will be labeled as -

Naive Bayes Classifier - Example (cont.)

Training phase outputs following probability estimates (tables bellow) used in **testing phase** to classify new examples.

$P(H \mid L)$	L = +	L = -
H = Yes	$\frac{1}{4}$	$\frac{2}{3}$
H = No	$\frac{3}{4}$	$\frac{1}{3}$

$P(S \mid L)$	L = +	L = -
S = L	$\frac{3}{4}$	$\frac{1}{3}$
S = M	$\frac{1}{4}$	$\frac{2}{3}$

$P(C \mid L)$	L = +	L = -
C = R	$\frac{1}{4}$	$\frac{1}{3}$
C = B	$\frac{1}{4}$	0
C = G	$\frac{2}{4}$	$\frac{2}{3}$

	L = +	L = -
$P(L)$	$\frac{4}{7}$	$\frac{3}{7}$

- Test2

$$\frac{P(L=+|H=Yes, S=M, C=B)}{P(L=-|H=Yes, S=M, C=B)}$$
$$= \frac{P(H=Yes|L=+)P(S=M|L=+)P(C=B|L=+)\cdot P(L=+)}{P(H=Yes|L=-)P(S=M|L=-)P(C=B|L=-)\cdot P(L=-)}$$
$$= \frac{Something}{0} > 1$$

$\Rightarrow$ Test2 will be labeled as +

- Issue: Observed fractions are poor estimates when we have small training set and large number of attribute values. When these are zero they will dominate the calculations for all test example described with that specific attribute value.

Naive Bayes Classifier - Example (cont.)

Training phase using Laplace smoothing outputs following probability estimates

$P(H\mid L)$	L = +	L = -
H = Yes	$\frac{1+1}{4+2}$	$\frac{2+1}{3+2}$
H = No	$\frac{3+1}{4+2}$	$\frac{1+1}{3+2}$

$P(S\mid L)$	L = +	L = -
S = L	$\frac{3+1}{4+2}$	$\frac{1+1}{3+2}$
S = M	$\frac{1+1}{4+2}$	$\frac{2+1}{3+2}$

$P(C\mid L)$	L = +	L = -
C = R	$\frac{1+1}{4+3}$	$\frac{1+1}{3+3}$
C = B	$\frac{1+1}{4+3}$	$\frac{0+1}{3+3}$
C = G	$\frac{2+1}{4+3}$	$\frac{2+1}{3+3}$

	L = +	L = -
$P(L)$	$\frac{4+1}{7+2}$	$\frac{3+1}{7+2}$

- Conditional probability estimates (no smoothing) $P(A_i = a_i|v_j) = \frac{n_{ij}}{n_j}$ where
  n_j is the number of training examples with class label v_j
  n_ij is the number of training examples with class label v_j and attribute value a_i
- Laplace smoothing $P(A_i = a_i|v_j) = \frac{n_{ij}+1}{n_j+m}$
  where m is the number of unique values attribute A_i can have.
- Test2 classification based on Laplace-smoothed estimates

$$\frac{P(L=+|H=Yes, S=M, C=B)}{P(L=-|H=Yes, S=M, C=B)}$$

$$= \frac{P(H=Yes|L=+)P(S=M|L=+)P(C=B|L=+)\cdot P(L=+)}{P(H=Yes|L=-)P(S=M|L=-)P(C=B|L=-)\cdot P(L=-)}$$

$$= \frac{375}{567} < 1 \Rightarrow \text{Test2 will be labeled as -}$$

Naive Bayes Classifier - Example (cont.)

Training phase using Laplace smoothing outputs following probability estimates

$P(H\|L)$	L = +	L = -
H = Yes	$\frac{1+1}{4+2}$	$\frac{2+1}{3+2}$
H = No	$\frac{3+1}{4+2}$	$\frac{1+1}{3+2}$

$P(S\|L)$	L = +	L = -
S = L	$\frac{3+1}{4+2}$	$\frac{1+1}{3+2}$
S = M	$\frac{1+1}{4+2}$	$\frac{2+1}{3+2}$

$P(C\|L)$	L = +	L = -
C = R	$\frac{1+1}{4+3}$	$\frac{1+1}{3+3}$
C = B	$\frac{1+1}{4+3}$	$\frac{0+1}{3+3}$
C = G	$\frac{2+1}{4+3}$	$\frac{2+1}{3+3}$

	L = +	L = -
$P(L)$	$\frac{4+1}{7+2}$	$\frac{3+1}{7+2}$

- Missing attribute values in test examples: just omit those attribute values in the calculations
- Test3 classification based on Laplace-smoothed estimates

$$\frac{P(L=+|H=Yes,S=M,C=?)}{P(L=-|H=Yes,S=M,C=?)}$$

$$=\frac{P(H=Yes|L=+)P(S=M|L=+)\cdot P(L=+)}{P(H=Yes|L=-)P(S=M|L=-)\cdot P(L=-)}$$

$< 1 \Rightarrow$ Test3 will be labeled as -

Naive Bayes for Document Classification

Problem statement

- Given a training data set of labeled documents and an unlabeled text document d and set of possible document classes C that d can be labeled with, we want to find the most probable class $c \in C$.

$$\begin{aligned} c_{MAP} &= \arg\max_{c \in C} P(c|d) \\ &= \arg\max_{c \in C} \frac{P(d|c)P(c)}{P(d)} \\ &= \arg\max_{c \in C} P(d|c)P(c) \end{aligned}$$

1. How should we represent an arbitrary document d in terms of attribute values?
2. How should we estimate the probabilities ($P(d|c)$, $P(c)$) required by Naive Bayes?

Document representation

- Document is a sequence of n words (including punctuation)
- n attribute values $< x_1, \ldots, x_i, \ldots, x_n >$, with the i-th attribute representing the i-th word position, and its value being the word appearing at the i-th position

  Example: "kiwi birds are native species in NZ."

  $< x_1=$"kiwi", $x_2=$birds",$x_3=$"are",$x_4=$"native",$x_5=$ "species",$x_6=$"in", $x_7=$"NZ", $x_8 =$ "." $>$
- Let X be the set of unique attribute values (e.g. unique words occurring in the training data set)

$$\begin{aligned} c_{MAP} &= \arg\max_{c \in C} P(d|c)P(c) \\ &= \arg\max_{c \in C} P(x_1, x_2, \ldots, x_n|c)P(c) \end{aligned}$$

Naive Bayes for Document Classification

- **Default assumption**: document attributes are independent given the document class, i.e., probability of words in one text position are independent of the words occurring in the other positions given the document class

$$\begin{aligned} c_{MAP} &= \arg\max_{c \in C} P(x_1|c) \cdot P(x_2|c) \cdot \ldots \cdot P(x_n|c)P(c) \\ &= \arg\max_{c \in C} P(c) \prod_i P(x_i|c) \end{aligned}$$

- We need to estimate $n \cdot |X| \cdot |C| + |C|$ probability terms.[1]
  Even for simple use cases this is prohibitively large, e.g., about 10 million terms for $|X| = 50000$, $|C| = 2$, $n = 100$.

[1] Since $\sum_{x \in X} P(x|c) = 1$ and $\sum_{c \in C} P(c) = 1$ we only need to estimate $n \cdot (|X| - 1) \cdot |C| + |C| - 1$ probability terms, also called the Naive Bayes model parameters

Naive Bayes for Document Classification

- **Additional assumption**: attributes are identically distributed given the document class, i.e., the probability of seeing a specific word w is independent of the specific word position in the document, that is

$$P(x_i = w|c) = P(x_j = w|c) \text{ for } 1 \leq i \neq j \leq n$$

- We can use the same training data set to get more reliable estimates, because we need to estimate only $|X| \cdot |C| + |C|$ probability terms.
- Probability estimates based on word counts in the training data set

$$P(x_i = w|c) = \frac{\overbrace{count(w, c)}^{\text{number of times the word occurs across all documents labled } c}}{\underbrace{\sum_{x \in X} count(x, c)}_{\text{effectively the total length of all documents labled } c}}$$

Estimating Probabilities – Laplace Smoothing

$$P(x_i = w|c) = \frac{count(w,c)}{\sum_{x \in X} count(x,c)}$$

- What if $count(w,c) = 0$?
- Solution: Simply add a constant

$$P(x_i = w|c) = \frac{count(w,c) + 1}{\sum_{x \in X}(count(x,c) + 1)} = \frac{count(w,c) + 1}{\sum_{x \in X} count(x,c) + |X|}$$

An Illustrative Example

	ID	Words	class
Training	d1	Kiwi, Sheep, Kiwi	NZ
	d2	Kiwi,Kiwi,Bird	NZ
	d3	Kiwi,Auckland	NZ
	d4	Munich,Oktoberfest,Kiwi	DE
Test	d5	Kiwi,Kiwi,Kiwi,Munich,Oktoberfest	?NZ

$C = \{NZ, DE\}$ $X = \{Kiwi, Sheep, Brid, Auckland, Munich, Oktoberfest\}$

$$P(c) = \frac{count(c)}{\text{number of docs}}$$

$$P(w|c) = \frac{count(w,c)+1}{\sum_{x \in X} count(x,c) + |X|}$$

- Priors

$$P(NZ) = 3/4$$
$$P(DE) = 1/4$$

- Posterior probabilities (to predict class of d5)

$$P(NZ|d5) \sim 3/4 * (3/7)^3 * 1/14 * 1/14 \approx 0.0003$$
$$P(DE|d5) \sim 1/4 * (2/9)^3 * 2/9 * 2/9 \approx 0.0001$$

- Conditional Probabilities

$$P(Kiwi|NZ) = (5+1)/(8+6) = 3/7$$
$$P(Munich|NZ) = (0+1)/(8+6) = 1/14$$
$$P(Oktoberfest|NZ) = (0+1)/(8+6) = 1/14$$
$$P(Kiwi|DE) = (1+1)/(3+6) = 2/9$$
$$P(Munich|DE) = (1+1)/(3+6) = 2/9$$
$$P(Oktoberfest|DE) = (1+1)/(3+6) = 2/9$$

Discussion

Are Naive Bayes Classier affected/robust to

- isolated noise examples?
- irrelevant attributes?
- missing values?
- correlated attributes?

Summary

- Bayes classifiers learn some target function $f : X \rightarrow Y$, or equivalently, $P(Y|X)$, by using the training data to learn estimates of $P(X|Y)$ and $P(Y)$
- New X examples can then be classified using these estimated probability distributions, plus Bayes' theorem.
- Issue: Typically requires an unrealistic number of training examples,i.e., more than the number of possible instances defined on the input (attribute) space.
- Solution: The Naive Bayes classifier assumes all attributes describing X are conditionally independent given Y: dramatically reduces the number of parameters that must be estimated to learn the classifier (meaning smaller training data sets are sufficient for the learning task).

Bayesian Networks

Bayesian Networks

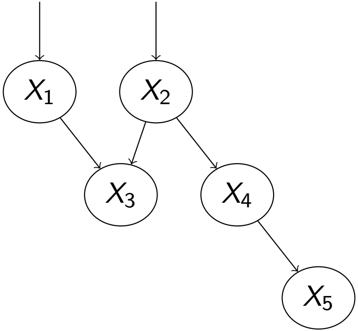
- Naive Bayes assumption of conditional independence too restrictive
- But it's intractable without such assumptions
- Bayesian Belief networks allow conditional independence among subsets of variables
- Allows combining prior knowledge about (in)dependencies among variables with observed training data

Bayesian Belief Networks

- A simple, graphical notation for conditional independence assertions and hence for compact specification of full joint distributions
- Syntax:
  - A set of nodes, one per random variable X_i
  - A set of directed edges between the nodes, each edge represents the direction of influence among random variables
  - A directed, acyclic graph (no cycles)
  - A conditional probability distribution for each node given its parent nodes: $P(X_i|Parents(X_i))$
- Conditional distribution represented as a conditional probability table (CPT) giving the distribution of X_i for each combination of parent values

Bayesian Belief Networks – Variable Relationships

- A graphical model of relationships
  - **Encodes dependencies among the variables** (**can** be causal dependencies)
    - X is a parent of Y, if there is a directed edge from X to Y.
    - X is an ancestor of Y (and Y is a descendant of X), if there is a directed path in the graph from X to Y. Example: X_1, X_2 are parents of $X3$, and X_2 is the parent of $X4$; X_2 has no ancestors/parents, but X_2 is ancestor of X_3, X_4 and X_5.
    - Local Markov property: A node is conditionally independent of its non-descendants given its parents. Example: X_5 is conditionally independent of X_2, X_3, X_1 given X_4; X_4 is conditionally independent of X_3, X_1 given X_2.



Bayesian Belief Networks – Variable Relationships

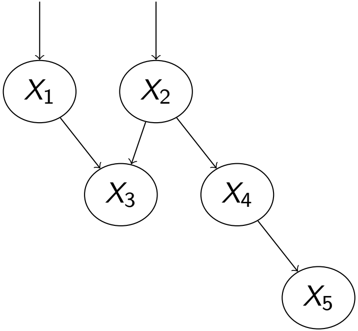
- A graphical model of relationships
  - **Encodes dependencies among the variables**
  - **Gives a specification of the joint probability distribution**

    If we label the nodes such that X_i is an ancestor of X_j only if $i < j$, the joint probability can be factorized using the chain rule of probability and local Markov property as

$$P(X_1, \ldots, X_n) = P(X_1)\underbrace{P(X_2|X1) \cdots P(X_n|X_1, X_2, \ldots, X_{n-1})}_{P(X_2,\ldots,X_n|X_1)}$$

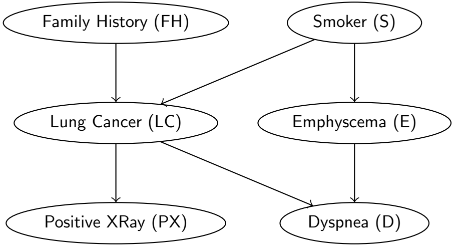
$$= \prod_{i=1}^{n} P(X_i|\underbrace{X_1, X_2, \ldots, X_{i-1}}_{\text{non-descendants of } X_i})$$

$$= \prod_{i=1}^{n} P(X_i|Parents(X_i))$$



Example: $P(X_1, X_2, X_3, X_4, X_5) = P(X_1)P(X_2)P(X3|X_1, X_2)P(X4|X2)P(X_5|X_4)$

Bayesian Belief Network – Example



Note[2]	$P(FH, S)$	$P(FH, \neg S)$	$P(\neg FH, S)$	$P(\neg FH, \neg S)$
LC	0.8	0.5	0.7	0.1
$\neg LC$	0.2	0.5	0.3	0.9

- The conditional probability table stored at the node *Lung Cancer*
- Shows the conditional probability for each possible combination of its parents (*Family History* and *Smoker*)
- Homework: Write the joint probability distribution for this network (in a factorized form)

[2] $\neg X$ for a Boolean variable X encodes $X =$ false, while $X =$ true is denoted simply as X

Bayesian (Belief) Network – Another Example

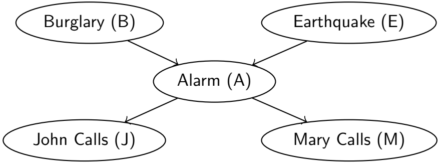
$P(B)$	$P(\neg B)$
0.001	0.999

$P(E)$	$P(\neg E)$
0.002	0.998

B	E	$P(A)$	$P(\neg A)$
t	t	0.95	0.05
t	f	0.94	0.06
f	t	0.29	0.71
f	f	0.001	0.999

A	$P(J)$	$P(\neg J)$
t	0.9	0.1
f	0.05	0.95

A	$P(M)$	$P(\neg M)$
t	0.7	0.3
f	0.01	0.99



- What is the probability that an alarm has sounded ($A =$ true), Mary calls ($M =$ true) and John calls ($J =$ true), and there was no earthquake ($E=$ false) and no burglary ($B =$ false) ?

$$P(x_1, \ldots, x_n) = P(\neg B, \neg E, A, J, M) = \prod_{Xi \in \{B,E,A,J,M\}} P(X_i = x_i | Parents(X_i))$$

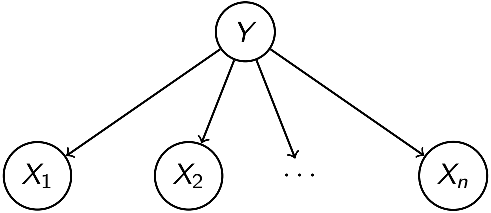
$$= P(\neg B) \cdot P(\neg E) \cdot P(A | \neg B \wedge \neg E) \cdot P(J|A) \cdot P(M|A)$$

$$= 0.999 \cdot 0.998 \cdot 0.001 \cdot 0.9 \cdot 0.7 = 0.00062$$

Application: Naive Bayes Classifier

Naive Bayes classifier is actually a special case of Bayesian network

- Variables: features X_i, class Y
- Graph structure:

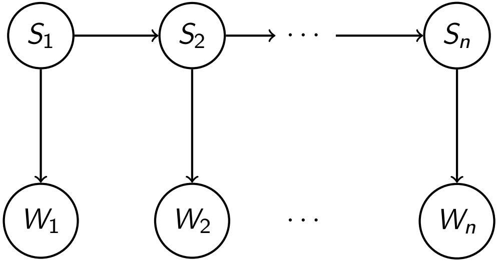


$$P(x_1, \dots, x_n, y) = P(y) \prod_{i=1}^{n} P(x_i | y)$$

Application: Hidden Markov Model

Consider a model that predicts tomorrow's weather given the weather condition so far.

- Variables: state S_i (hidden), weather W_i (observed)
- Graph structure:



$$
\begin{aligned}
&P(s_1, \dots, s_n, w_1, \dots, w_n) \\
&= P(s_1)P(w_1|s_1)P(s_2|s_1)P(w_2|s_2)\dots P(s_n|s_{n-1})P(w_n|s_n) \\
&= P(s_1)P(s_2|s_1)\dots P(s_n|s_{n-1}) \times P(w_1|s_1)P(w_2|s_2)\dots P(w_n|s_n)
\end{aligned}
$$

Inference in Bayesian Networks

- How can one infer the (probabilities of) values of one or more network variables, given observed values of others?
- Bayesian Networks contain all information needed for this inference
- If only one variable with unknown value (e.g. target in classification), easy to infer it, e.g., MAP
- In general case, the exact inference is NP-hard, approximations have been introduced (e.g. Monte Carlo-based methods and variational inference)
- In some cases, the structure will need to be learned from the training data as well.
- More on Bayesian Networks in Bishop's *Pattern Recognition And Machine Learning*

Summary

- Bayes optimal classifier combines the predictions of all alternative hypothesis weighted by their posterior probabilities
- Bayesian networks provide a natural representation for conditional independence
- Naive Bayes classifier a simple and fast method for classification that assumes attribute values are conditionally independent given the target value.

Literature

- Chapter 6 of Mitchell's *Machine Learning* (also look at Section 2 of `www.cs.cmu.edu/~tom/mlbook/NBayesLogReg.pdf`)
- Chapter 8 of Bishop's *Pattern Recognition and Machine Learning*

THE UNIVERSITY OF AUCKLAND
Te Whare Wānanga o Tāmaki Makaurau
NEW ZEALAND

SCIENCE
SCHOOL OF COMPUTER SCIENCE

Thank you for your attention!

`https://ml.auckland.ac.nz`