Bayesian Models and Logistic Regression

Probability theory as basis for the construction of classifiers:

- Multivariate probabilistic models
- Independence assumptions
- Naive Bayesian classifier
- Forest-augmented networks (FANs)
- Approximation: logistic regression

Notation

- Random (= statistical = stochastic) variable: upper-case letter, e.g. V or X, or upper-case string, e.g. RAIN
- *Binary* variables: take one of *two* values, $X =$ *true* (abbreviated x) and $X =$ *false* (abbreviated $\neg x$)
- Conjunctions: $(X = x) \wedge (Y = y)$ as $X = x, Y = y$
- Templates: X, Y means $X = x, Y = y$, for any value x, y, i.e. the choice of the values x and y does not really matter
- $\sum_X P(X) = P(x) + P(\neg x)$, where X is binary

Joint probability distribution

Joint (= multivariate) distribution:

$P(X_1, X_2, \ldots, X_n)$

Example of joint probability distribution:

$$
\begin{array}{rcl}
P(X_1, X_2, X_3) & \text{with:} & \\
P(x_1, x_2, x_3) & = & 0.1 \\
P(\neg x_1, x_2, x_3) & = & 0.05 \\
P(x_1, \neg x_2, x_3) & = & 0.10 \\
P(x_1, x_2, \neg x_3) & = & 0.0 \\
P(\neg x_1, \neg x_2, x_3) & = & 0.3 \\
P(x_1, \neg x_2, \neg x_3) & = & 0.2 \\
P(\neg x_1, x_2, \neg x_3) & = & 0.1 \\
P(\neg x_1, \neg x_2, \neg x_3) & = & 0.15
\end{array}
$$

Note that: $\sum_{X_1, X_2, X_3} P(X_1, X_2, X_3) = 1$

Marginalisation:

$$P(x_3) = \sum_{X_1, X_2} P(X_1, X_2, x_3) = 0.55$$

Chain rule

Definition of conditional probability distribution:

$$P(X_1 \mid X_2, \ldots, X_n) = \frac{P(X_1, X_2, \ldots, X_n)}{P(X_2, \ldots, X_n)}$$

$$\Rightarrow P(X_1, X_2, \ldots, X_n) = P(X_1 \mid X_2, \ldots, X_n) P(X_2, \ldots, X_n)$$

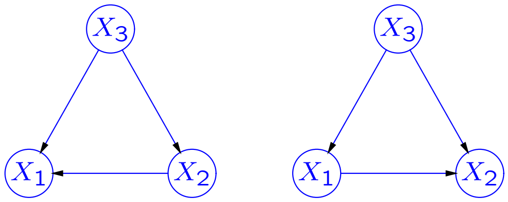
Furthermore,

$$
\begin{array}{rcl}
P(X_2, \ldots, X_n) & = & P(X_2 \mid X_3, \ldots, X_n) P(X_3, \ldots, X_n) \\
\vdots & \vdots & \vdots \\
P(X_{n-1}, X_n) & = & P(X_{n-1} \mid X_n) P(X_n) \\
P(X_n) & = & P(X_n)
\end{array}
$$

Chain rule yields factorisation:

$$P\left(\bigwedge_{i=1}^{n} X_i\right) = \prod_{i=1}^{n} P\left(X_i \mid \bigwedge_{k=i+1}^{n} X_k\right)$$

Chain rule - digraph



(1) (2)

Factorisation (1):

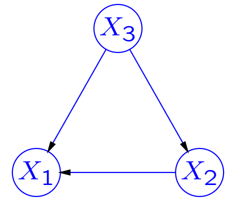
$$P(X_1, X_2, X_3) = P(X_1 \mid X_2, X_3) \cdot P(X_2 \mid X_3) P(X_3)$$

Other factorisation (2):

$$P(X_1, X_2, X_3) = P(X_2 \mid X_1, X_3) \cdot P(X_1 \mid X_3) P(X_3)$$

$\Rightarrow$ different *factorisations* possible

Does the chain rule help?



$$P(X_1, X_2, X_3) = P(X_1 \mid X_2, X_3) \cdot P(X_2 \mid X_3) P(X_3)$$

i.e. we need:

$$P(x_1 \mid x_2, x_3)$$
$$P(x_1 \mid \neg x_2, x_3)$$
$$P(x_1 \mid x_2, \neg x_3)$$
$$P(x_1 \mid \neg x_2, \neg x_3)$$
$$P(x_2 \mid x_3)$$
$$P(x_2 \mid \neg x_3)$$
$$P(x_3)$$

Note $P(\neg x_1 \mid x_2, x_3) = 1 - P(x_1 \mid x_2, x_3)$, etc.
$\Rightarrow$ 7 probabilities required (as for $P(X_1, X_2, X_3)$)

Use stochastic independence

$$P(X_1, X_2, X_3) = P(X_2 \mid X_1, X_3) \cdot P(X_3 \mid X_1) P(X_1)$$

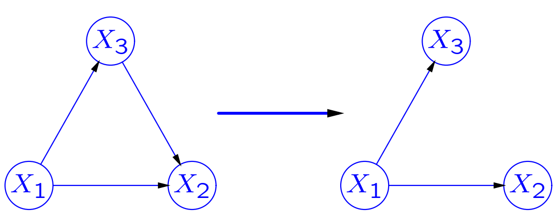
Assume that X_2 and X_3 are conditionally independent given X_1:

$$P(X_2 \mid X_1, X_3) = P(X_2 \mid X_1)$$

and

$$P(X_3 \mid X_1, X_2) = P(X_3 \mid X_1)$$

Notation: $X_2 \perp\!\!\!\perp X_3 \mid X_1$, $X_3 \perp\!\!\!\perp X_2 \mid X_1$



Only $5 = 2 + 2 + 1$ probabilities (instead of 7)

Definition Bayesian Network (BN)

A Bayesian network $\mathcal{B}$ is a pair $\mathcal{B} = (G, P)$, where:

- $G = (V(G), A(G))$ is an acyclic directed graph, with
  - $V(G) = \{X_1, X_2, \ldots, X_n\}$, a set of vertices (nodes)
  - $A(G) \subseteq V(G) \times V(G)$ a set of arcs
- $P : \wp(V(G)) \to [0, 1]$ is a joint probability distribution, such that

  $$P(V(G)) = \prod_{i=1}^{n} P(X_i \mid \pi_G(X_i))$$

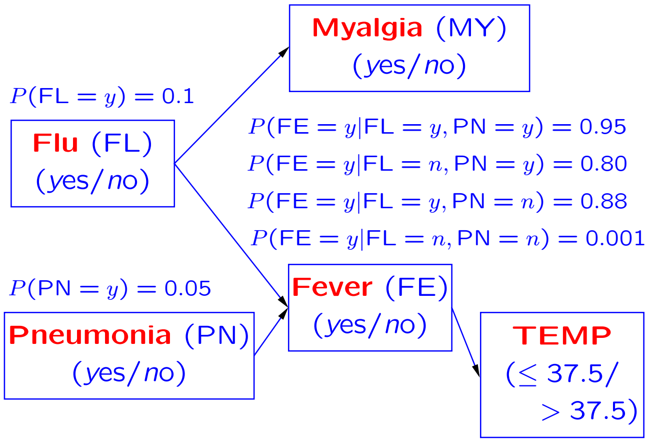
  where $\pi_G(X_i)$ denotes the set of immediate ancestors (parents) of vertex X_i in G

Example Bayesian network

$$\boxed{P(\text{FL}, \text{PN}, \text{MY}, \text{FE}, \text{TEMP})}$$

$P(\text{MY} = y | \text{FL} = y) = 0.96$

$P(\text{MY} = y | \text{FL} = n) = 0.20$

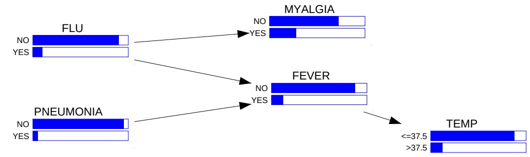


$P(\text{TEMP} \leq 37.5 | \text{FE} = y) = 0.1$

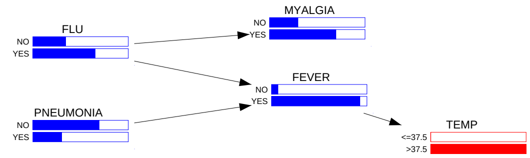
$P(\text{TEMP} \leq 37.5 | \text{FE} = n) = 0.99$

Reasoning: evidence propagation

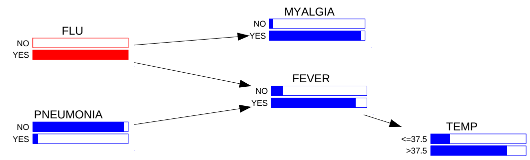
- Nothing known:



- Temperature >37.5 °C:



- Likely symptoms of the flu?

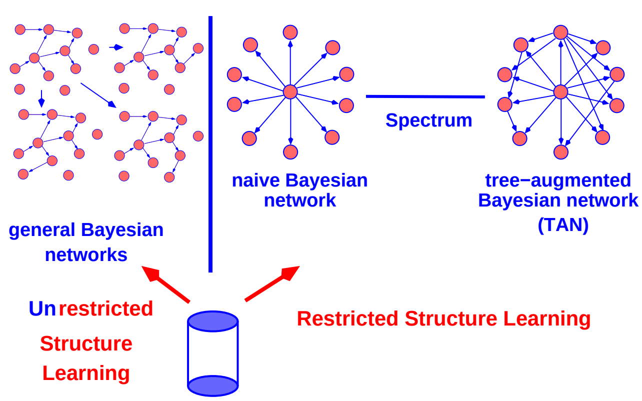


Bayesian network structure learning

Bayesian network $\mathcal{B} = (G, P)$, with

- digraph $G = (V(G), A(G))$, and
- probability distribution

$$P(V) = \prod_{X \in V(G)} P(X \mid \pi(X))$$



Special form Bayesian networks

Problems:

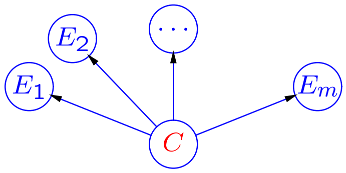
- for many BNs too many probabilities have to be assessed
- complex BNs do not necessarily yield better classifiers
- complex BNs may yield better estimates to the probability distribution

Solution:

use simple probabilistic models for classification:

- *naive* (independent) form BN
- *T*ree-*A*ugmented Bayesian *N*etwork (TAN)
- *F*orest-*A*ugmented Bayesian *N*etwork (FAN)

Naive (independent) form BN



- C is a **class variable**
- The **evidence variables** E_i in the evidence $\mathcal{E} \subseteq \{E_1, \ldots, E_m\}$ are conditionally independent given the class variable C

This yields, using Bayes' rule:

$$P(C \mid \mathcal{E}) = \frac{P(\mathcal{E} \mid C)P(C)}{P(\mathcal{E})}$$

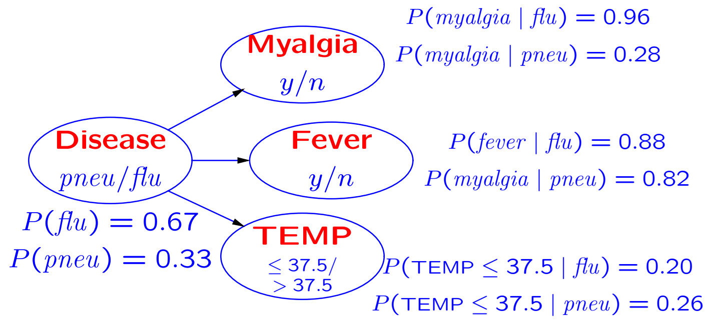
with, as $E_i \perp\!\!\!\perp E_j \mid C$, for $i \neq j$:

$$P(\mathcal{E} \mid C) = \prod_{E \in \mathcal{E}} P(E \mid C) \qquad \text{by cond. ind.}$$

$$P(\mathcal{E}) = \sum_{C} P(\mathcal{E} \mid C)P(C) \qquad \text{marg. \& cond.}$$

Classifier: $c_{\max} = \arg\max_C P(C \mid \mathcal{E})$

Example of naive Bayes



Evidence: $\mathcal{E} = \{\text{TEMP} > 37.5\}$; computation of the probability of flu using Bayes' rule:

$$P(flu \mid \text{TEMP} > 37.5) = \frac{P(\text{TEMP} > 37.5) \mid flu)P(flu)}{P(\text{TEMP} > 37.5)}$$

$$P(\text{TEMP} > 37.5) = P(\text{TEMP} > 37.5 \mid flu)P(flu) + P(\text{TEMP} > 37.5 \mid pneu)P(pneu) = 0.8 \cdot 0.67 + 0.74 \cdot 0.33 \approx 0.78$$

$$\Rightarrow P(flu \mid \text{TEMP} \geq 37.5) = 0.8 \cdot 0.67/0.78 \approx 0.687$$

Computing probabilities from data

Compute the weighted average of

- estimate $\widehat{P}_D(V \mid \pi(V))$ of the conditional probability distribution for variable V based on the dataset D
- Dirichlet prior Θ, which reflects prior knowledge

These are combined as follows:

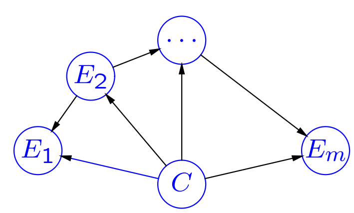
$$P_D(V \mid \pi(V)) = \frac{n}{n+n_0}\widehat{P}_D(V \mid \pi(V)) + \frac{n_0}{n+n_0}\Theta$$

where

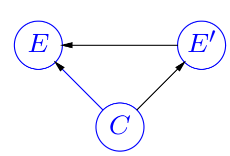
- n is the size of the dataset D
- n_0 is the estimated size of the (virtual) 'dataset' on which the prior knowledge is based (*equivalence sample size*)

More complex Bayesian networks

- We want to add arcs to a naive Bayesian network to improve its performance
- Result: possibly TAN



- Which arc should be added?

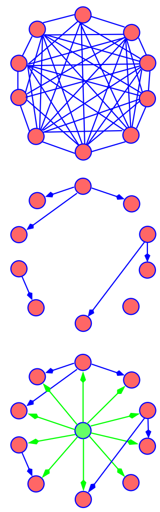


Compute mutual information between variables E, E' *conditioned* on the class variable C:

$$I_P(E, E' \mid C) = \sum_{E,E',C} P(E, E', C) \cdot \log \frac{P(E, E' \mid C)}{P(E \mid C)P(E' \mid C)}$$

FAN algorithm

Choose $k \geq 0$. Given evidence variables E_i, a class variable C, and a dataset D:



1. Compute mutual information $-I_P(E_i, E_j \mid C)$ $\forall (E_i, E_j)$, $i \neq j$, in a complete undirected graph
2. Construct a minimum-cost spanning forest containing exactly k edges
3. Change each tree in the forest into a directed tree
4. Add an arc from the class vertex C to every evidence vertex E_i in the forest
5. Learn conditional probability distributions from D using Dirichlet distributions

Performance evaluation

- Success rate σ based on:

$$c_{max} = \text{argmax}_c \{P(c \mid \mathbf{x}_i)\}$$

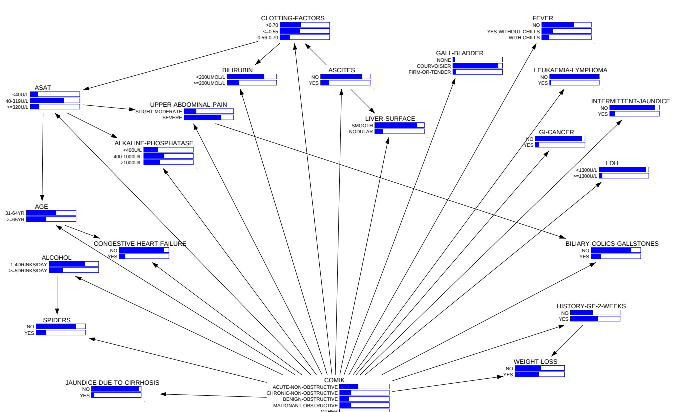
for $x_i \in D$, $i = 1, \ldots, n = |D|$

- Total entropy (penalty):

$$E = -\sum_{i=1}^{n} \ln P(c \mid \mathbf{x}_i)$$

  - if $P(c \mid \mathbf{x}_i) = 1$, then $\ln P(c \mid \mathbf{x}_i) = 0$
  - if $P(c \mid \mathbf{x}_i) \downarrow 0$ then $\ln P(c \mid \mathbf{x}_i) \to -\infty$

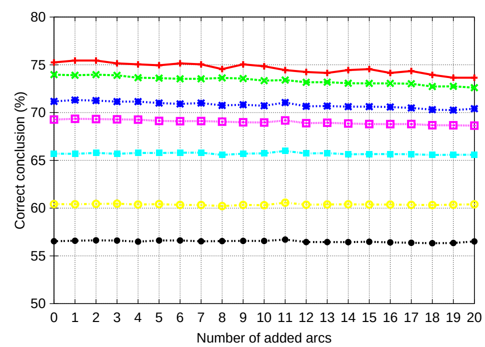
Example COMIK BN

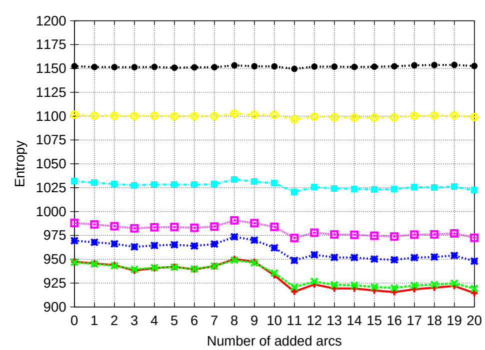


Based on COMIK dataset:

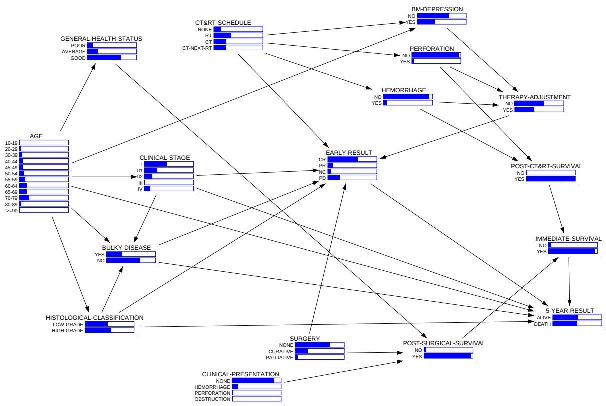
- Dataset with 1002 patient cases with liver and biliary tract disease, collected by the Danish COMIK group
- Has been used as vehicle for learning various probabilistic models

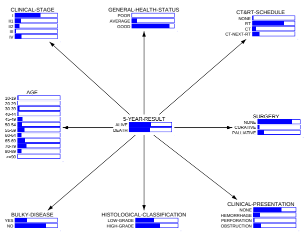
Results for COMIK Dataset

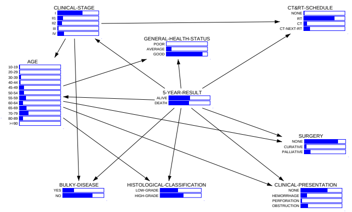




Comparison Bayesian networks: NHL







Naive Bayes: odds-likelihood form

For class variable C and evidence $\mathcal{E}$:

$$P(C \mid \mathcal{E}) = \frac{\prod_{E \in \mathcal{E}} P(E \mid C) P(C)}{P(\mathcal{E})}$$

if $E \perp\!\!\!\perp E' \mid C$, $\forall E, E' \in \mathcal{E}$; for $C = \mathit{true}$:

$$P(c \mid \mathcal{E}) = \frac{\prod_{E \in \mathcal{E}} P(E \mid c) P(c)}{P(\mathcal{E})}$$

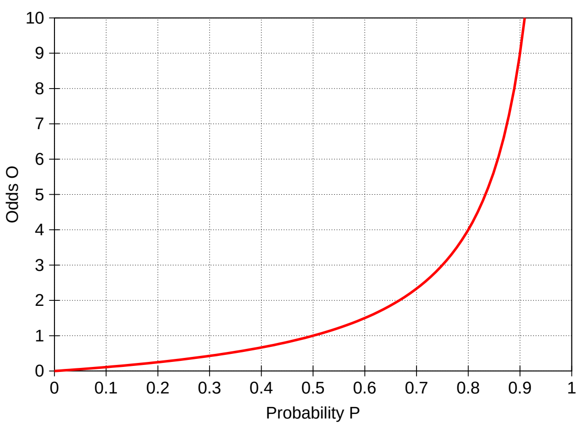
For $C = \mathit{false}$:

$$P(\neg c \mid \mathcal{E}) = \frac{\prod_{E \in \mathcal{E}} P(E \mid \neg c) P(\neg c)}{P(\mathcal{E})}$$

$$\begin{aligned}
\Rightarrow \frac{P(c \mid \mathcal{E})}{P(\neg c \mid \mathcal{E})} &= \frac{\prod_{E \in \mathcal{E}} P(E \mid c)}{\prod_{E \in \mathcal{E}} P(E \mid \neg c)} \frac{P(c)}{P(\neg c)} \\
&= \prod_{i=1}^{m} \lambda_i \cdot O(c) \\
&= O(c \mid \mathcal{E})
\end{aligned}$$

Here is $O(c \mid \mathcal{E})$ the conditional odds, and $\lambda_i = P(E_i \mid c)/P(E_i \mid \neg c)$ is a likelihood ratio

Odds and probabilities



Note that:

- $O(c \mid \mathcal{E}) = 1$ if $P(c \mid \mathcal{E}) = 0.5$
- $O(c \mid \mathcal{E}) \to \infty$ if $P(c \mid \mathcal{E}) \uparrow 1$

Odds, likelihoods and logarithms

Odds:

$$\begin{aligned}
O(c \mid \mathcal{E}) &= \frac{P(c \mid \mathcal{E})}{P(\neg c \mid \mathcal{E})} \\
&= \frac{P(c \mid \mathcal{E})}{1 - P(c \mid \mathcal{E})}
\end{aligned}$$

Back to probabilities:

$$P(c \mid \mathcal{E}) = \frac{O(c \mid \mathcal{E})}{1 + O(c \mid \mathcal{E})}$$

Logarithmic odds-likelihood form:

$$\begin{aligned}
\ln O(c \mid \mathcal{E}) &= \ln \prod_{i=1}^{m} \lambda_i \cdot O(c) \\
&= \sum_{i=1}^{m} \ln \lambda_i + \ln O(c) \\
&= \sum_{i=0}^{m} \omega_i
\end{aligned}$$

with $\omega_0 = \ln O(c)$ and $\omega_i = \ln \lambda_i$, $i = 1, \ldots, m$

Log-odds and weights

Log-odds:

$$\begin{aligned}\ln O(c \mid \mathcal{E}) &= \ln \prod_{i=1}^{m} \lambda_i \cdot O(c)\\ &= \sum_{i=1}^{m} \ln \lambda_i + \ln O(c)\\ &= \sum_{i=0}^{m} \omega_i\end{aligned}$$

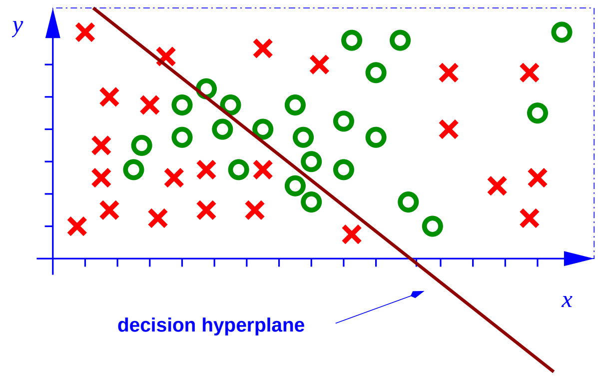
Back to probabilities:

$$\begin{aligned}P(c \mid \mathcal{E}) &= \frac{O(c \mid \mathcal{E})}{1 + O(c \mid \mathcal{E})}\\ &= \frac{\exp(\sum_{i=0}^{m} \omega_i)}{1 + \exp(\sum_{i=0}^{m} \omega_i)}\end{aligned}$$

Adjust ω_i with weights β_i based in existing interactions between variables:

$$\ln O(c \mid \mathcal{E}) = \sum_{i=0}^{m} \beta_i \omega_i = \beta^T \omega$$

Logistic regression



Hyperplane: $\{\omega \mid \beta^T \omega = 0\}$ where

- $c = \beta_0 \omega_0$ is the intercept (recall that $\omega_0 = \ln O(c)$, which is independent of any evidence E)
- ω_i, $i = 1, \ldots, m$ correspond to the probabilities we want to find

Maximum likelihood estimate

Database D, $|D| = N$, with independent instances $\mathbf{x}_i \in D$, then likelihood function l:

$$l(\beta) = \prod_{i=1}^{N} P_\beta(C_i \mid \mathbf{x}_i')$$

where C_i is the class value for instance i, and $\mathbf{x}_i'$ is $\mathbf{x}_i$, without C_i

Log-likelihood function L:

$$\begin{aligned}L(\beta) &= \ln l(\beta)\\ &= \sum_{i=1}^{N} \ln P_\beta(C_i \mid \mathbf{x}_i')\\ &= \sum_{i=1}^{N} \Big(y_i \ln P_\beta(c_i \mid \mathbf{x}_i') +\\ &\qquad (1 - y_i) \ln(1 - P_\beta(c_i \mid \mathbf{x}_i'))\Big)\end{aligned}$$

where c_i is (always) the value $C_i = \mathit{true}$; $y_i = 1$ if c_i is the class value for x_i, $y_i = 0$, otherwise

Maximum likelihood estimate

$$\begin{aligned}L(\beta) &= \sum_{i=1}^{N} \Big(y_i \ln P_\beta(c_i|\mathbf{x}_i') + (1 - y_i) \ln(1 - P_\beta(c_i|\mathbf{x}_i'))\Big)\\ &= \sum_{i=1}^{N} \Big(y_i \beta^T \omega - \ln\Big(1 + e^{\beta^T \omega}\Big)\Big)\end{aligned}$$

Maximisation of $L(\beta)$:

$$\begin{aligned}\frac{\partial L(\beta)}{\partial \beta} &= \sum_{i=1}^{N} \left(y_i \omega - \frac{e^{\beta^T \omega}}{1 + e^{\beta^T \omega}}\right)\\ &= \sum_{i=1}^{N} \left(y_i \omega - P_\beta(c_i \mid \mathbf{x}_i')\right)\\ &= 0\end{aligned}$$

Can be solved using equation solving methods, such as Newton-Raphson method

$$\beta_{r+1} = \beta_r - \frac{\partial L(\beta)}{\partial \beta} \cdot \left(\frac{\partial^2 L(\beta)}{\partial \beta^T \partial \beta}\right)^{-1}$$

for $r = 0, 1, \ldots$, and $\beta_0 = 0$; result: approximation for β

WEKA output

```
Scheme:       weka.classifiers.Logistic
Relation:     weather.symbolic
Instances:    14
Attributes:   5
              outlook
              temperature
              humidity
              windy
              play
Test mode:    evaluate on training data

=== Classifier model (full training set) ===

Logistic Regression (2 classes)

Coefficients...
Variable      Coeff.
       1     34.9227
       2    -48.1161
       3      7.8472
       4     17.3933
       5    -33.0445
       6     22.2601
       7    -82.415
       8    -54.6671
Intercept    66.1064
```