PDSS

Overview

- Decision support
  - Markov blanket
  - Utility
  - Optimal decision
  - Sequential decision
    - Optimal stopping
    - Value of information
  - Examples for optimal decision
  - Risk models and their characterization

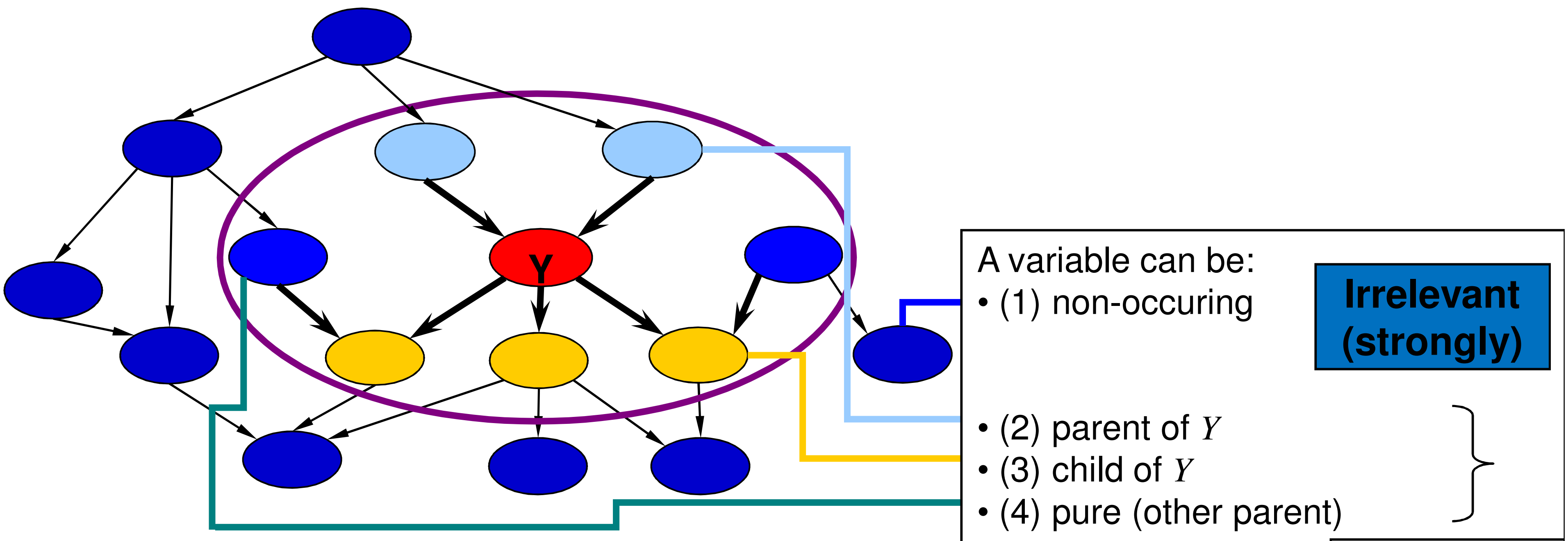
Inference in Bayesian networks

- (Passive, observational) inference
  - P(Query|Observations)
- Interventionist inference
  - P(Query|Observations, Interventions)
- Counterfactual inference
  - P(Query| Observations, Counterfactual conditionals)

- Biomedical applications
  - Prevention
  - Screening
  - Diagnosis
  - Therapy selection
  - Therapy modification

The Markov Blanket

A minimal sufficient set for prediction/diagnosis.



Markov Blanket Sets (MBS) the set of nodes which probabilistically isolate the target from the rest of the model

Markov Blanket Membership (MBM)

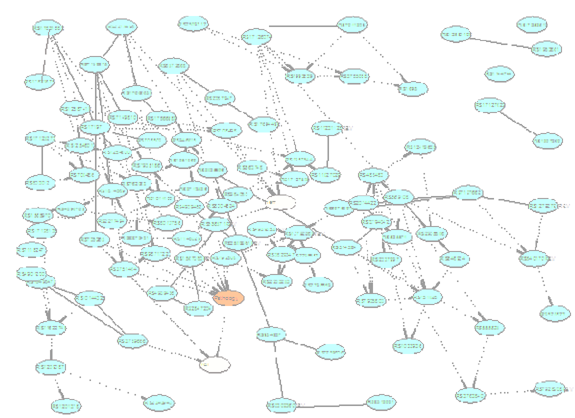
(symmetric) pairwise relationship induced by MBS

Bayes-omics

- **Thomas Bayes (c. 1702 – 1761)**
- Bayesian probability
- Bayes' rule $p(Cause \mid Effect) \propto p(Effect \mid Cause) \times p(Cause)$
- Bayesian statistics $p(\text{Model} \mid \text{Data}) \propto p(\text{Data} \mid \text{Model})\, p(\text{Model})$
- **<u>Bayesian decision</u>** $a^* = \arg\max_i \sum_j U(o_j)\, p(o_j \mid a_i)$
- Bayesian model averaging $p(prediction \mid data) =$

$$= \sum_i p(pred. \mid Model_i)\, p(Model_i \mid data)$$

- Bayesian networks
- Bayes factor
- Bayes error
- Bayesian „communication"
- ...

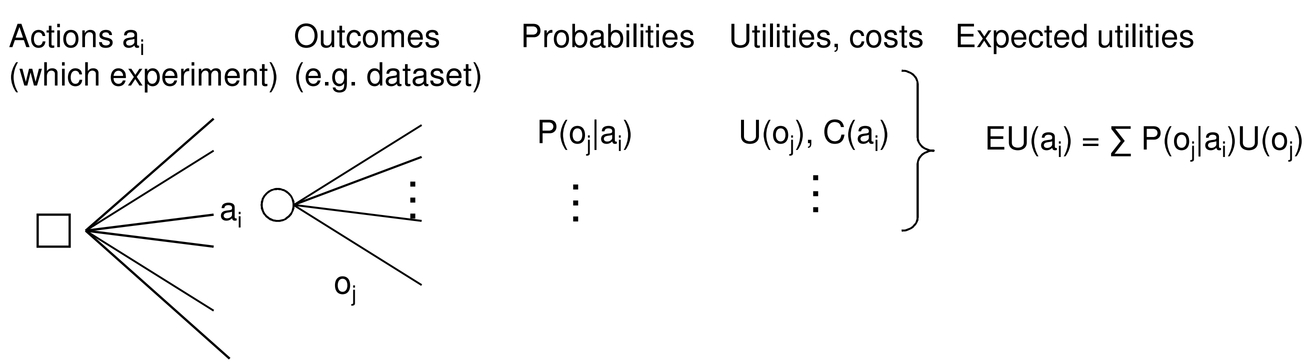


Decision theory
probability theory+utility theory

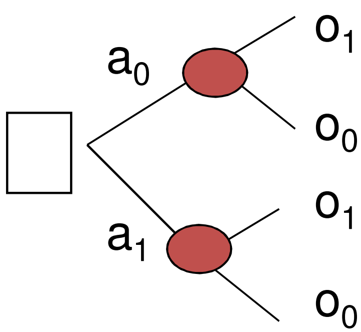
- Decision situation:
  - Actions a_i
  - Outcomes o_j
  - Probabilities of outcomes $p(o_j \mid a_i)$
  - Utilities/losses of outcomes $U(o_j \mid a_i)$
    - QALY, micromort
  - Maximum Expected Utility Principle (MEU)
    - Best action is the one with maximum expected utility

$$EU(a_i) = \sum_j U(o_j \mid a_i)\, p(o_j \mid a_i)$$

$$a^* = \arg\max_i EU(a_i)$$



Optimal binary decision



reported	Ref.:0	Ref.1
0	$C_{0\|0}$	$C_{0\|1}$
1	$C_{1\|0}$	$C_{1\|1}$

If the outcome y and the prediction $\hat{y}$ are binary, the loss is defined by a binary cost matrix $C_{\hat{y}|y}$. The minimal loss decision is defined by

$$\arg\min_{\hat{y}} C_{\hat{y}|0} P(Y=0|\boldsymbol{x}) + C_{\hat{y}|1} P(Y=1|\boldsymbol{x}), \tag{8}$$

In case of $C_{0|0} = C_{1|1} = 0$, the prediction $\hat{y} = 1$ is optimal if

$$\tau = \frac{C_{1|0}}{C_{1|0} + C_{0|1}} \leq P(Y=1|\boldsymbol{x}) \tag{9}$$

where $\tau \in [0, 1]$ is the optimal decision threshold.

Frequentist vs Bayesian decision theory

- Bayesian decision theory:
  - Probabilities of outcomes
  - Utilities of outcomes
  - Expected Utility Principle
- Classical decision theory:
  - Neyman-Pearson
  - „Hippocratic Oath”(?)

reported	Ref.:0	Ref.1
0	$C_{0\|0}$	$C_{0\|1}$
1	$C_{1\|0}$	$C_{1\|1}$

reported	Ref.: 0	Ref.1
0	TN	FN
1	FP	TP

reported	Ref.0/null	Ref.:1
0		Type II
1	Type I („false rejection”)	

Utilities

Utilities map states to real numbers. Which numbers?

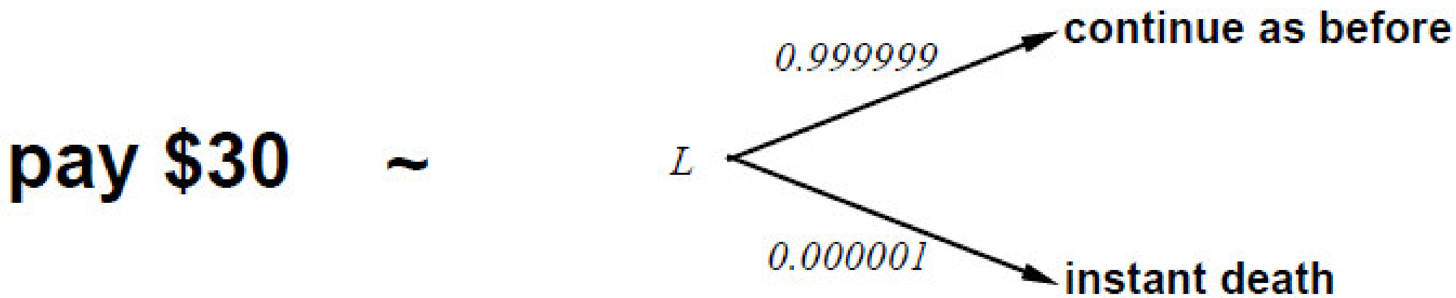
Standard approach to assessment of human utilities:

compare a given state A to a standard lottery L_p that has

"best possible prize" $u_\top$ with probability p

"worst possible catastrophe" $u_\bot$ with probability $(1-p)$

adjust lottery probability p until $A \sim L_p$



Russel&Norvig: Artificial intelligence, ch.16

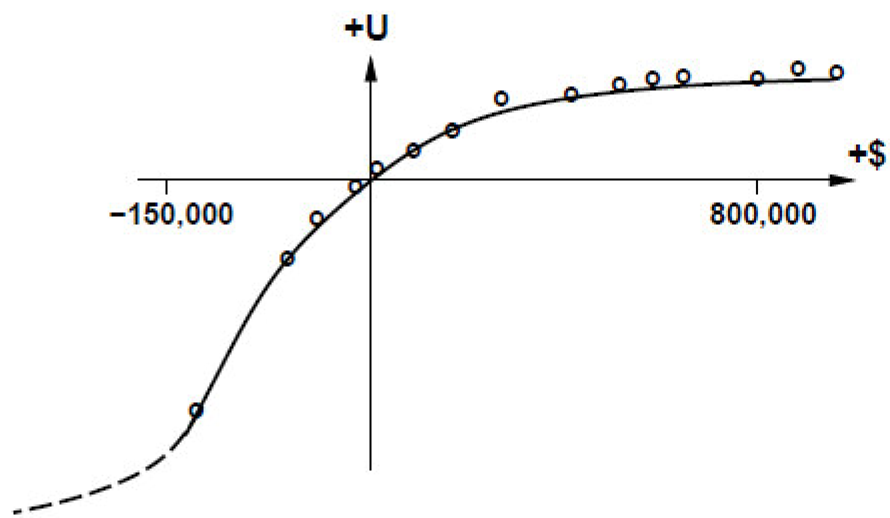
Utility of money

Money does **not** behave as a utility function

Given a lottery L with expected monetary value $EMV(L)$,
usually $U(L) < U(EMV(L))$, i.e., people are risk-averse

Utility curve: for what probability p am I indifferent between a prize x and a lottery $[p, \$M; \ (1-p), \$0]$ for large M?

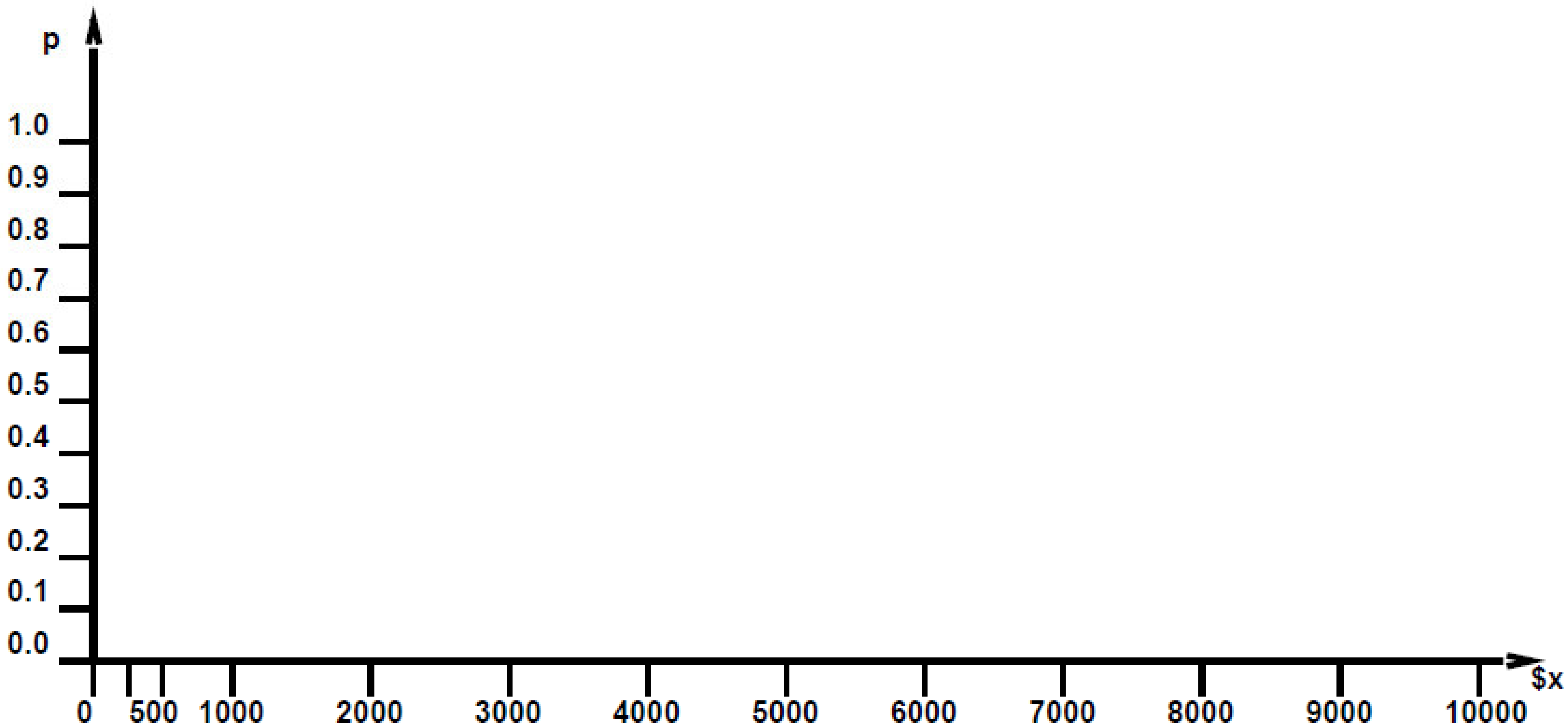
Typical empirical data, extrapolated with risk-prone behavior:



Russel&Norvig: Artificial intelligence, ch.16

Estimating the utility of money

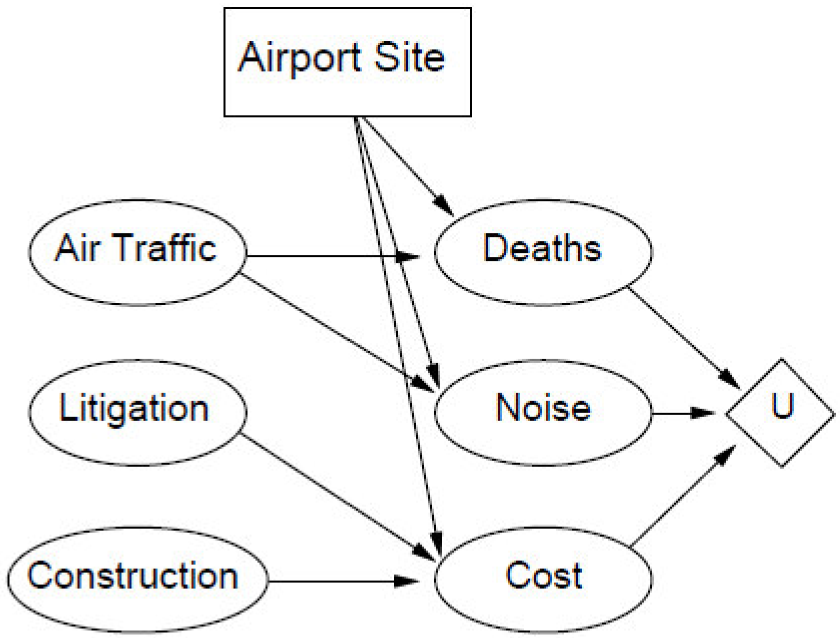
For each x, adjust p until half the class votes for lottery (M=10,000)



Russel&Norvig: Artificial intelligence, ch.16

Decision networks

Add action nodes and utility nodes to belief networks
to enable rational decision making



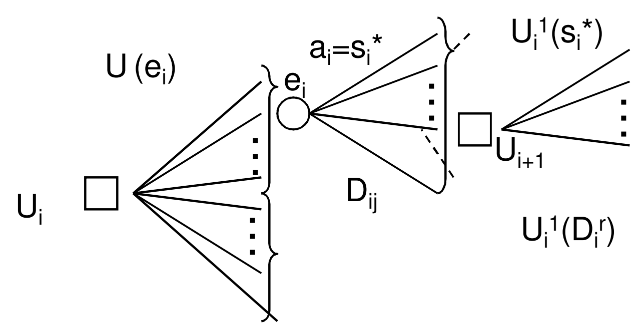
Algorithm:

For each value of action node
    compute expected value of utility node given action, evidence
Return MEU action

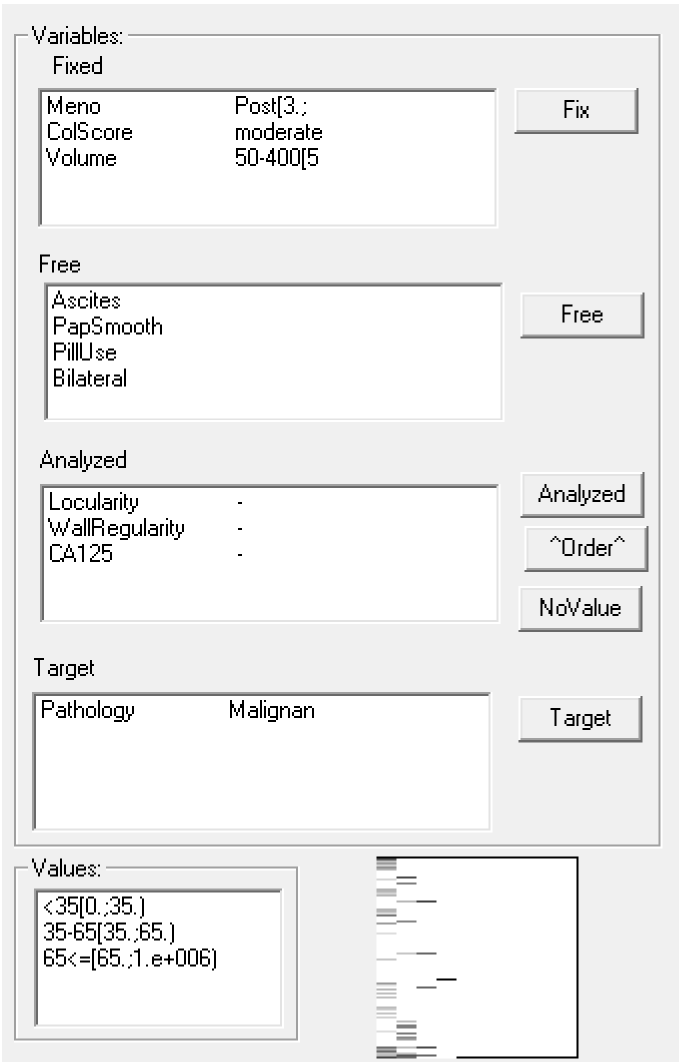
Russel&Norvig: Artificial intelligence, ch.16

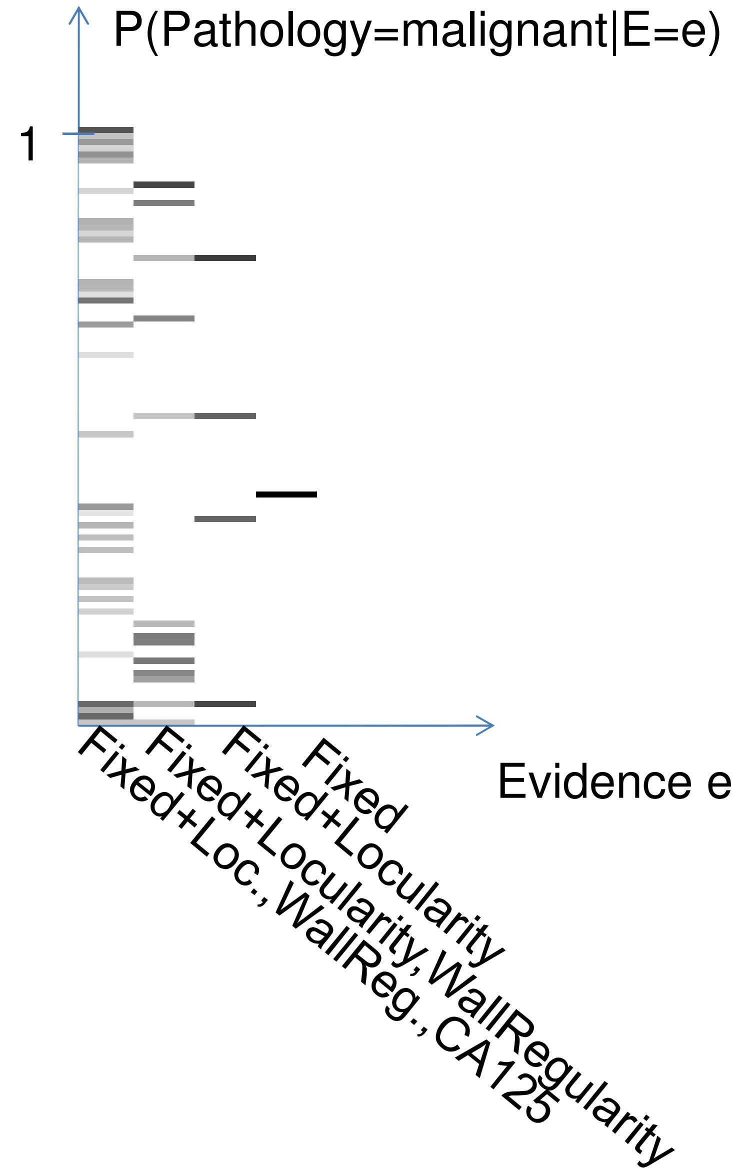
Extensions

- Bayesian learning
  - Predictive inference
  - Parametric inference
- Value of further information
- Sequential decisions
  - Optimal stopping (secretary problem)
  - Multiarmed bandit problem
  - Markov decision problem
  - ….



Sensitivity of the inference





Value of (perfect) information: Vo(P)I

Current evidence E, current best action α
Possible action outcomes S_i, potential new evidence E_j

$$EU(\alpha|E) = \max_a \textstyle\sum_i U(S_i)\ P(S_i|E,a)$$

Suppose we knew $E_j = e_{jk}$, then we would choose $\alpha_{e_{jk}}$ s.t.

$$EU(\alpha_{e_{jk}}|E, E_j = e_{jk}) = \max_a \textstyle\sum_i U(S_i)\ P(S_i|E, a, E_j = e_{jk})$$

E_j is a random variable whose value is *currently* unknown
$\Rightarrow$ must compute expected gain over all possible values:

$$VPI_E(E_j) = \left(\textstyle\sum_k P(E_j = e_{jk}|E) EU(\alpha_{e_{jk}}|E, E_j = e_{jk})\right) - EU(\alpha|E)$$

(VPI = value of perfect information)

Russel&Norvig: Artificial intelligence, ch.16

Properties of VoPI

Nonnegative—in **expectation**, not **post hoc**

$$\forall j, E \;\; VPI_E(E_j) \geq 0$$

Nonadditive—consider, e.g., obtaining E_j twice

$$VPI_E(E_j, E_k) \neq VPI_E(E_j) + VPI_E(E_k)$$

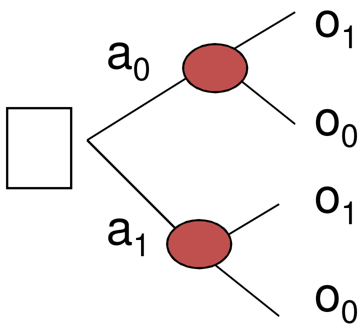
Order-independent

$$VPI_E(E_j, E_k) = VPI_E(E_j) + VPI_{E,E_j}(E_k) = VPI_E(E_k) + VPI_{E,E_k}(E_j)$$

Note: when more than one piece of evidence can be gathered,
maximizing VPI for each to select one is not always optimal
$\Rightarrow$ evidence-gathering becomes a **sequential** decision problem

Russel&Norvig: Artificial intelligence, ch.16

Example: preoperative diagnosis (evidence-based medicine)



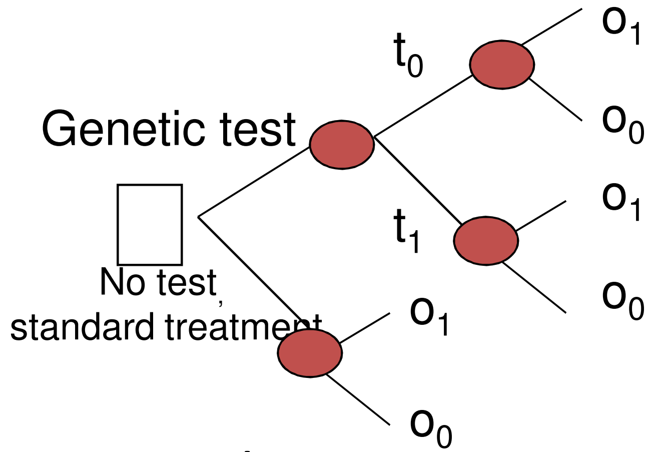
reported	Ref.:0	Ref.1
0	$C_{0\|0}$	$C_{0\|1}$
1	$C_{1\|0}$	$C_{1\|1}$

- Assume
  - Correct decision has no penalty: $C_{0|0}=C_{1|1}=0$
  - FalsePositive decision causes a modest loss: $C_{1|0}$=10000$
  - FalseNegative decision causes a heavy loss: $C_{0|1}$=90000$
- If our belief is $p(Y=1|X=x)=p$, then
  - Expected loss of decision 0 is $pC_{0|1}$
  - Expected loss of decision 1 is $(1-p)\ C_{1|0}$

➔Decision 1 is optimal if its loss is smaller: $pC_{0|1} > (1-p)\ C_{1|0}$

then $p > C_{0|1}/(C_{0|1}+C_{1|0})$, i.e. if $p > 0.1$

Example: personalized treatment



reported	Ref.:0	Ref.1
0	0	$C_{0\|1}$
1	$C_{1\|0}$	0

- Assume that genetic test t
  - has cost C_t
  - two outcomes t_0, t_1 with probability $p(t_1)=q$
  - can be used in treatment selection $p(Y=1|X=x, t_i)=p_i$
- The value of the test is: $EL - (1-q)EL_0 + qEL_1$
  - Expected loss without the test is: $EL=\min(pC_{0|1},(1-p)\,C_{1|0})$
  - Expected loss with the test is $(1-q)EL_0 + qEL_1$
    - t_0: $EL_0=\min(p_0C_{0|1},(1-p_0)\,C_{1|0})$
    - t_1: $EL_1=\min(p_1C_{0|1},(1-p_1)\,C_{1|0})$

$\Rightarrow$ If $EL_0 \approx EL$, then $(1-q)EL_0 + qEL_1 - EL \approx q(EL_1 - EL)$, e.g. $q(p-p_1)C_{0|1}$

Risk models

- Multivariate methods
  - Linear models $Y = \sum_{i=0}^{n} \beta_i I_j x_i$
  - Logistic regression, decision trees, kernel methods,..

Logistic regression (LR): $P(y|\underline{x}) = \sigma[\sum_{i=0}^{n}(\beta_i x_i + \sum_{j=1}^{n} (\beta_{i,j} x_i x_j + \ldots)))]$,

Multilayer perceptron (MLPs): $f(\underline{x}, \underline{\omega}) = \sigma[\sum_{i=1}^{L}(\omega_i \ \tanh[\sum_{j=1}^{|\underline{X}|}(\omega_{ij} x_j + \omega_{i0})])]$,

Naive Bayesian networks (N-BNs): $p(y, x_1, \ldots, x_n|\underline{\theta}) = p(y) \prod_{i=1}^{n} p(x_i|y)$,

Bayesian networks (BNs): $p(x_1, \ldots, x_n|\underline{\theta}, G) = \prod_{i=1}^{n} p(x_i|\, \mathrm{pa}(X_i, G))$.

Logistic regression

Recall: NaiveBN!

Assume binary outcomes $y, \bar{y}$ and predictors $x_i, \bar{x}_i$. Logistic regression without interactions can be defined by the odds ratios for the predictors x_i, $i = 1, \ldots, n$ and the bias Ψ_0 ($x_0 \triangleq 1$):

$$\Psi_i = \frac{P(y|x_i)P(\bar{y}|\bar{x}_i)}{P(\bar{y}|x_i)P(y|\bar{x}_i)} \triangleq \exp^{\beta_i}, \Psi_0 = \prod_{i=0}^{n} \frac{P(y|\bar{x}_i)}{P(\bar{y}|\bar{x}_i)} \triangleq \exp^{\beta_0} .$$

The odds $P(y|\boldsymbol{x})/P(\bar{y}|\boldsymbol{x})$ for a given $\boldsymbol{x}$ is defined as

$$P(y|\boldsymbol{x})/P(\bar{y}|\boldsymbol{x}) = \prod_{i=0}^{n} \Psi_i^{x_i} \qquad (18)$$

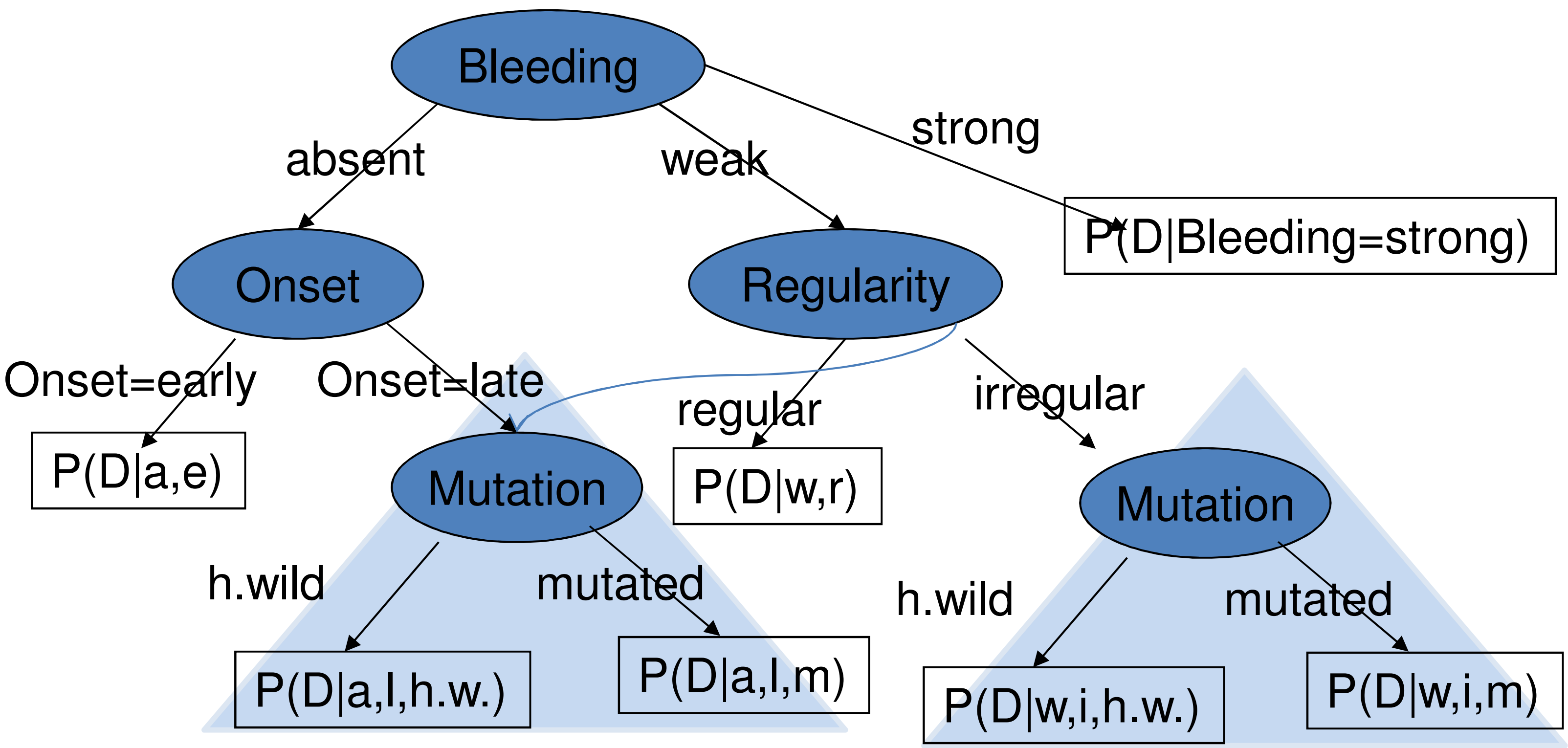
$$\log(P(y|\boldsymbol{x})/P(\bar{y}|\boldsymbol{x})) = \sum_{i=0}^{n} \beta_i x_i \qquad (19)$$

$$P(y|\boldsymbol{x}) = \sigma(\sum_{i=0}^{n} \beta_i x_i), \qquad (20)$$

where $\sigma()$ is the logistic sigmoid function$\sigma(x) = 1/(1 + e^{-x})$.

$$P(y|\boldsymbol{x}) = \sigma[\sum_{i=0}^{n}(\beta_i x_i + \sum_{j=1}^{n}(\beta_{i,j} x_i x_j + \sum_{k=1}^{n}(\beta_{i,j,k} x_i x_j x_k + \ldots)))],$$

Decision trees, decision graphs



Decision tree: Each internal node represent a (univariate) test, the leafs contains the conditional probabilities given the values along the path.
Decision graph: If conditions are equivalent, then subtrees can be merged.
E.g. If (Bleeding=absent,Onset=late) ~ (Bleeding=weak,Regularity=irreg)

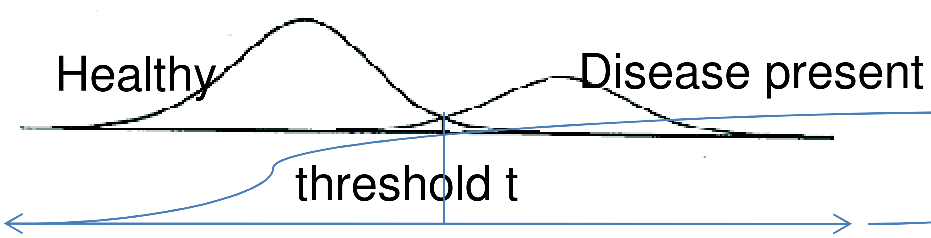
Characterizing a decision function

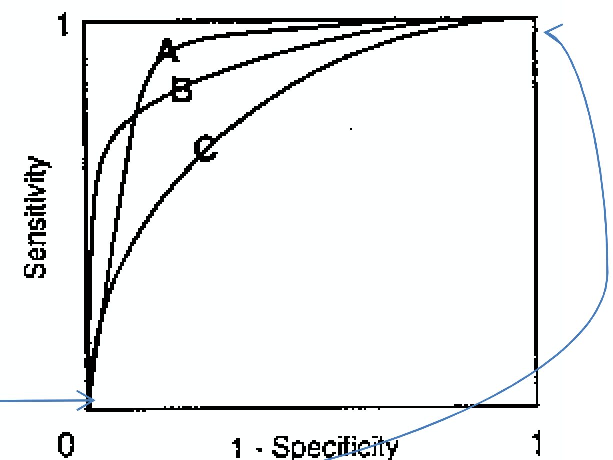
Goal: selection of a decision function $g : R^d \to 0, 1$.

1. Sensitivity = TP/(TP +FN), Specificity = TN/(TN + FP)
2. Positive predictive value = TP/(TP + FP), Negative predictive value = TN/(TN + FN)
3. Misclassification rate = (FP + FN)/(TP + FP + FN + TN)

If decision function g is defined by a scalar function $f(x) : R^d \to R$ and threshold t that $f(x) : 0$, if $g(x) < t$, 1 otherwise, then we can compute the Area Under the Receiver Operating Characteristics Curve (ROC,AUC). AUC is the probability that two random samples from class 0 and 1 is correctly classified.

Sensitivity: p(Prediction=TRUE|Ref=TRUE)
Specificity: p(Prediction=FALSE|Ref=FALSE)
PPV: p(Ref=TRUE|Prediction=TRUE)
NPV: p(Ref=FALSE|Prediction=FALSE)





Summary

- Decision support
  - Markov blanket
  - Utility
  - Optimal decision
  - Sequential decision
    - Optimal stopping
    - Value of information
  - Risk models
  - Measuring the quality of a decision function