T-61.3050 Machine Learning: Basic Principles

Bayesian Networks

Kai Puolamäki

Laboratory of Computer and Information Science (CIS)
Department of Computer Science and Engineering
Helsinki University of Technology (TKK)

Autumn 2007

Outline

Rules of Probability

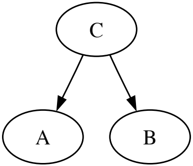
- $P(E, F) = P(F, E)$: probability of both E and F happening.
- $P(E) = \sum_F P(E, F)$ (sum rule, marginalization)
- $P(E, F) = P(F \mid E)P(E)$ (product rule, conditional probability)
- Consequence: $P(F \mid E) = P(E \mid F)P(F)/P(E)$ (Bayes' formula)
- We say E and F are independent if $P(E, F) = P(E)P(F)$ (for all E and F).
- We say E and F are conditionally independent given G if $P(E, F \mid G) = P(E \mid G)P(F \mid G)$, or equivalently $P(E \mid F, G) = P(E \mid G)$.

Bayesian Networks

Bayesian network is a directed acyclic graph (DAG) that describes a joint distribution over the vertices $X_1,\ldots,X_d$ such that

$$P(X_1,\ldots,X_d) = \prod_{i=1}^{d} P(X_i \mid \mathrm{parents}(X_i)),$$

where $\mathrm{parents}(X_i)$ are the set of vertices from which there is an edge to X_i.



$$P(A,B,C) = P(A \mid C)P(B \mid C)P(C).$$

(A and B are conditionally independent given C.)

Outline

Inference in Bayesian Networks

- When structure of the Bayesian network and the probability factors are known, one usually wants to do inference by computing conditional probabilities.
- This can be done with the help of the sum and product rules.
- Example: probability of the cat being on roof if it is cloudy, $P(F \mid C)$?

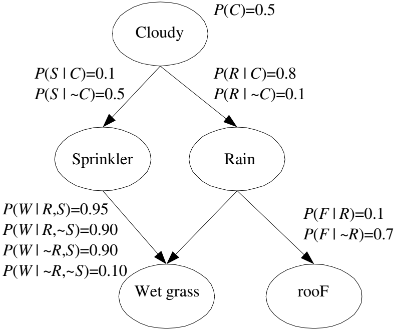
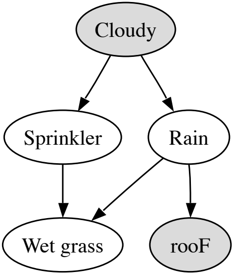


Figure 3.5 of Alpaydin (2004).

Inference in Bayesian Networks

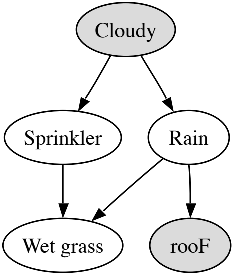
- Example: probability of the cat being on roof if it is cloudy, $P(F \mid C)$?
- S, R and W are unknown or hidden variables.
- F and C are observed variables. Conventionally, we denote the observed variables by gray nodes (see figure on the right).
- We use the product rule $P(F \mid C) = P(F, C)/P(C)$, where $P(C) = \sum_F P(F, C)$.
- We must sum over or marginalize over hidden variables S, R and W: $P(F, C) = \sum_S \sum_R \sum_W P(C, S, R, W, F)$.



$P(C, S, R, W, F) = P(F \mid R)P(W \mid S, R)P(S \mid C)P(R \mid C)P(C)$

Inference in Bayesian Networks

$$P(F,C) =$$
$$P(C,S,R,W,F) + P(C,-S,R,W,F)$$
$$+P(C,S,-R,W,F) + P(C,-S,-R,W,F)$$
$$+P(C,S,R,-W,F) + P(C,-S,R,-W,F)$$
$$+P(C,S,-R,-W,F) + P(C,-S,-R,-W,F)$$

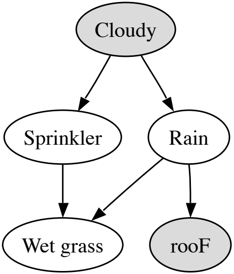


$$P(C,S,R,W,F) = P(F \mid R)P(W \mid S,R)P(S \mid C)P(R \mid C)P(C)$$

- We obtain similar formula for $P(F,-C)$, $P(-F,C)$ and $P(-F,-C)$.
- Notice: we have used shorthand F to denote $F=1$ and $-F$ to denote $F=0$.
- In principle, we know the numeric value of each joint distribution, hence we can compute the probabilities.

Inference in Bayesian Networks

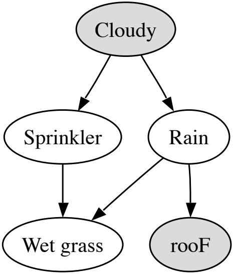
- There are 2^5 terms in the sums.
- Generally: marginalization is NP-hard, the most staightforward approach would involve a computation of $O(2^d)$ terms.
- We can often do better by smartly re-arranging the sums and products. Behold:
- Do the marginalization over W first: $P(C,S,R,F) = \sum_W P(F \mid R)P(W \mid S,R)P(S \mid C)P(R \mid C)P(C) = P(F \mid R)\sum_W [P(W \mid S,R)]P(S \mid C)P(R \mid C)P(C) = P(F \mid R)P(S \mid C)P(R \mid C)P(C)$.



$$P(C,S,R,W,F) = P(F \mid R)P(W \mid S,R)P(S \mid C)P(R \mid C)P(C)$$

Inference in Bayesian Networks

- Now we can marginalize over S easily: $P(C,R,F) = \sum_S P(F \mid R)P(S \mid C)P(R \mid C)P(C) = P(F \mid R)\sum_S [P(S \mid C)]P(R \mid C)P(C) = P(F \mid R)P(R \mid C)P(C)$.
- We must still marginalize over R: $P(C,F) = P(F \mid R)P(R \mid C)P(C) + P(F \mid -R)P(-R \mid C)P(C) = 0.1 \times 0.8 \times 0.5 + 0.7 \times 0.2 \times 0.5 = 0.11$.
- $P(C,-F) = P(-F \mid R)P(R \mid C)P(C) + P(-F \mid -R)P(-R \mid C)P(C) = 0.9 \times 0.8 \times 0.5 + 0.3 \times 0.2 \times 0.5 = 0.39$.
- $P(C) = P(C,F) + P(C,-F) = 0.5$.
- $P(F \mid C) = P(C,F)/P(C) = 0.22$.
- $P(-F \mid C) = P(C,-F)/P(C) = 0.78$.



$$P(C,S,R,W,F) = P(F \mid R)P(W \mid S,R)P(S \mid C)P(R \mid C)P(C)$$

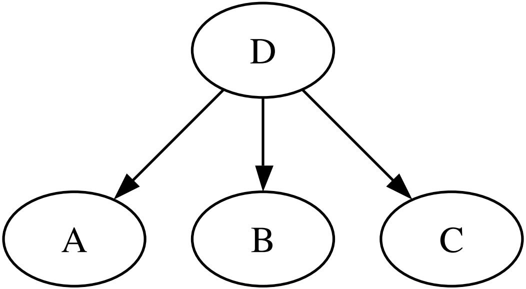
Bayesian Networks: Inference

- To do inference in Bayesian networks one has to marginalize over variables.
- For example: $P(X_1) = \sum_{X_2} \cdots \sum_{X_d} P(X_1, \ldots, X_d)$.
- If we have Boolean arguments the sum has $O(2^d)$ terms. This is inefficient!
- Generally, marginalization is a NP-hard problem.
- If Bayesian Network is a tree: Sum-Product Algorithm (a special case being Belief Propagation).
- If Bayesian Network is "close" to a tree: Junction Tree Algorithm.
- Otherwise: approximate methods (variational approximation, MCMC etc.)

Sum-Product Algorithm

- Idea: sum of products is difficult to compute. Product of sums is easy to compute, if sums have been re-arranged smartly.
- Example: disconnected Bayesian network with d vertices, computing $P(X_1)$.
  - sum of products: $P(X_1) = \sum_{X_2} \ldots \sum_{X_d} P(X_1) \ldots P(X_d)$.
  - product of sums:
    $P(X_1) = P(X_1) \left(\sum_{X_2} P(X_2)\right) \ldots \left(\sum_{X_d} P(X_d)\right) = P(X_1)$.
- Sum-Product Algorithm works if the Bayesian Network is directed tree.
- For details, see e.g., Bishop (2006).

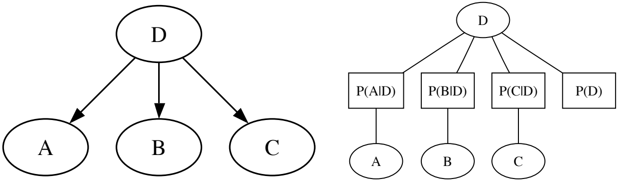
Sum-Product Algorithm
Example



$$P(A, B, C, D) = P(A \mid D)P(B \mid D)P(C \mid D)P(D)$$

Task: compute $\tilde{P}(D) = \sum_A \sum_B \sum_C P(A, B, C, D)$.

Sum-Product Algorithm
Example



$$P(A, B, C, D) = P(A \mid D)P(B \mid D)P(C \mid D)P(D)$$

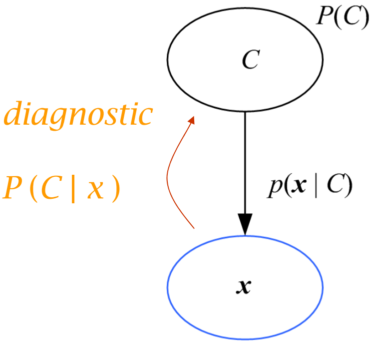
- Factor graph is composed of vertices (ellipses) and factors (squares), describing the factors of the joint probability.
- The Sum-Product Algorithm re-arranges the product (check!):

$$\begin{aligned} \tilde{P}(D) &= \left(\sum_A P(A \mid D)\right)\left(\sum_B P(B \mid D)\right)\left(\sum_C P(C \mid D)\right) P(D) \\ &= \sum_A \sum_B \sum_C P(A, B, C, D). \end{aligned} \quad (1)$$

Observations

- Bayesian network forms a partial order of the vertices. To find (one) total ordering of vertices: remove a vertex with no outgoing edges (zero out-degree) from the network and output the vertex. Iterate until the network is empty. (This way you can also check that the network is DAG.)
- If all variables are Boolean, storing a full Bayesian network of d vertices — or full joint distribution — as a look-up table takes $O(2^d)$ bytes.
- If the highest number of incoming edges (in-degree) is k, then storing a Bayesian network of d vertices as a look-up table takes $O(d2^k)$ bytes.
- When computing marginals, disconnected parts of the network do not contribute.
- Conditional independence is “easy” to see.

Bayesian Network: Classification



Bayes' rule inverts the arc:

$$P(C \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid C)P(C)}{p(\mathbf{x})}$$

Alpaydin (2004) Ch 3 / slides

Naive Bayes' Classifier

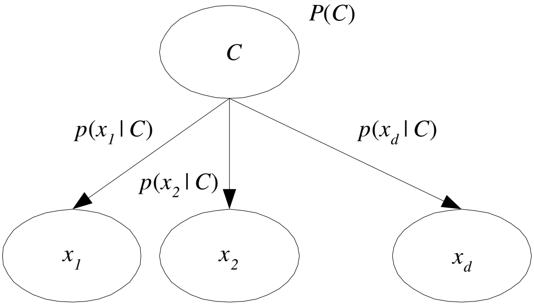
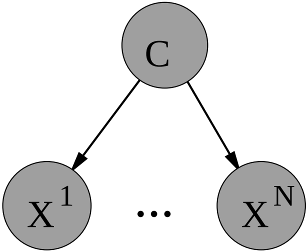


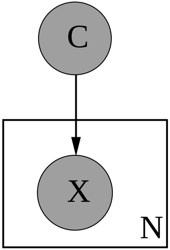
Figure 3.7 Alpaydin (2004).

- X^i are conditionally independent given C.
- $P(\mathcal{X}, C) = P(x^1 \mid C)P(x^2 \mid C) \ldots P(x^d \mid C)P(C)$.

Naive Bayes' Classifier



Equivalently:



- Plate is used as a shorthand notation for repetition. The number of repetitions is in the bottom right corner.
- Gray nodes denote observed variables.

Outline

1. Bayesian Networks
   - Reminders
   - Inference
   - Finding the Structure of the Network
2. Probabilistic Inference
   - Bernoulli Process
   - Posterior Probabilities
3. Estimating Parameters
   - Estimates from Posterior
   - Bias and Variance
   - Conclusion

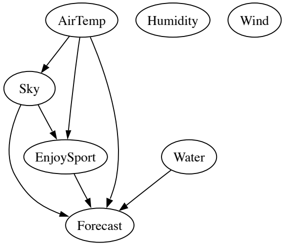
Finding the Structure of the Network

- Often, the network structure is given by an expert.
- In probabilistic modeling, the network structure defines the structure of the model.
- Finding an optimal Bayesian network structure is NP-hard
- Idea: Go through all possible network structures M and compute the likelihood of data $\mathcal{X}$ given the network structure $P(\mathcal{X} \mid M)$.
- Choose the network complexity appropriately.
- Choose network that, for a given network complexity, gives the best likelihood.
- The Bayesian approach: choose structure M that maximizes $P(M \mid \mathcal{X}) \propto P(\mathcal{X} \mid M)P(M)$, where $P(M)$ is a prior probability for network structure M (more complex networks should have smaller prior probability).

Finding a Network

- Full Bayesian network of d vertices and $d(d-1)/2$ edges describes the training set fully and the test set probably poorly.
- As before, in finding the network structure, we must control the complexity so that the the model generalizes.
- Usually one must resort to approximate solutions to find the network structure (e.g., DEAL package in R).
- A feasible exact algorithm exists for up to $d = 32$ variables, with a running time of $o(d^2 2^{d-2})$.
- See Silander et al. (2006) A Simple Optimal Approach for Finding the Globally Optimal Bayesian Network Structure. In Proc 22nd UAI. (pdf)

Finding a Network



Network found by Bene at `http://b-course.hiit.fi/bene`

t	*Sky*	*AirTemp*	*Humidity*	*Wind*	*Water*	*Forecast*	*EnjoySport*
1	Sunny	Warm	Normal	Strong	Warm	Same	1
2	Sunny	Warm	High	Strong	Warm	Same	1
3	Rainy	Cold	High	Strong	Warm	Change	0
4	Sunny	Warm	High	Strong	Cool	Change	1

Outline

Boys or Girls?

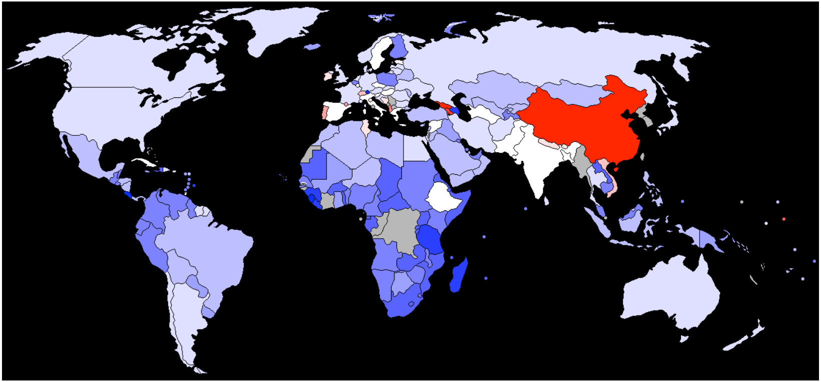


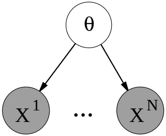
Figure: Sex ratio by country population aged below 15. Blue represents more women, red more men than the world average of 1.06 males/female. Image from Wikimedia Commons, author Dbachmann, GFDLv1.2.

Bernoulli Process

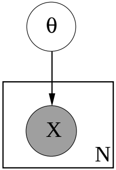
- The world average probability that a newborn child is a boy ($X = 1$) is about $\theta = 0.512$ [probability of a girl ($X = 0$) is then $1 - \theta = 0.488$].
- Bernoulli process:

$$P(X = x \mid \theta) = \theta^x (1 - \theta)^{1-x} \quad , \quad x \in \{0, 1\}.$$

- Assume we observe the genders of N newborn children, $\mathcal{X} = \{x^t\}_{t=1}^N$. What is the sex ratio?
- Joint distribution: $P(x^1, \ldots, x^N, \theta) = P(x^1 \mid \theta) \ldots P(x^N \mid \theta) P(\theta)$.
- Notice we must fix some prior for θ, $P(\theta)$.



Equivalently:



Outline

Comparing Models

- The likelihood ratio (Bayes factor) is defined by

$$BF(\theta_2;\theta_1) = \frac{P(\mathcal{X} \mid \theta_2)}{P(\mathcal{X} \mid \theta_1)}$$

- If we believe before seeing any data that the probability of model θ_1 is $P(\theta_1)$ and of model θ_2 is $P(\theta_2)$ then the ratio of their posterior probabilities is given by

$$\frac{P(\theta_2 \mid \mathcal{X})}{P(\theta_1 \mid \mathcal{X})} = \frac{P(\theta_2)}{P(\theta_1)} \times BF(\theta_1;\theta_2)$$

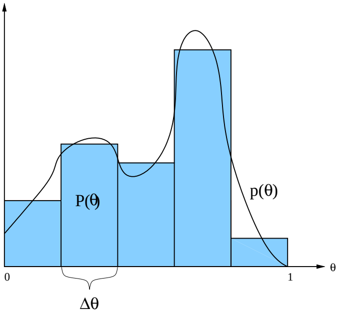
- This ratio allows us to compare our degrees of beliefs into two models.
- Posterior probability density allows us to compare our degrees of beliefs between infinite number of models after observing the data.

Discrete vs. Continuous Random Variables

- The Bernoulli parameter θ is a real number in $[0, 1]$.
- Previously we considered binary $(0/1)$ random variables.
- Generalization to multinomial random variables that can have values $1, 2, \ldots, K$ is straightforward.
- Generalization to continuous random variable: divide the interval $[0, 1]$ to K equally sized intervals of width $\Delta\theta = 1/K$. Define probability density $p(\theta)$ such that the probability of θ being in interval $S_i = [(i-1)\Delta\theta, i\Delta\theta]$, $i \in \{1, \ldots, K\}$, is $P(\theta \in S_i) = p(\theta')\Delta\theta$, where θ' is some point in S_i.
- At limit $\Delta\theta \to 0$:

$$E_{P(\theta)}\left[f(\theta)\right] = \sum_{\theta} P(\theta) f(\theta) \longrightarrow E_{p(\theta)}\left[f(\theta)\right] = \int d\theta p(\theta) f(\theta).$$

Discrete vs. Continuous Random Variables



- $P(\theta \in [(i-1)\Delta\theta, i\Delta\theta]) = p(\theta')\Delta\theta$.
- At limit $\Delta\theta \to 0$:

$$E_{P(\theta)}[f(\theta)] = \sum_{\theta} P(\theta)f(\theta) \longrightarrow E_{p(\theta)}[f(\theta)] = \int d\theta p(\theta)f(\theta).$$

Estimating the Sex Ratio

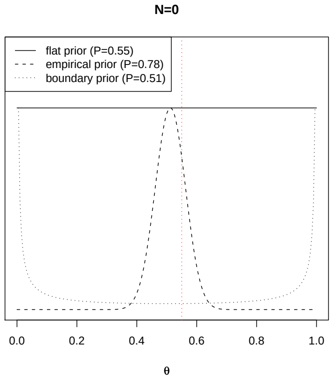
- Task: estimate the Bernoulli parameter θ, given N observations of the genders of newborns in an unnamed country.
- Assume the "true" Bernoulli parameter to be estimated in the unnamed country is $\theta = 0.55$, the global average being 51.2%.
- Posterior probability density after seeing N newborns in $\mathcal{X} = \{x^t\}_{t=1}^N$:

$$\begin{aligned} p(\theta \mid \mathcal{X}) &= \frac{p(\mathcal{X} \mid \theta) p(\theta)}{p(\mathcal{X})} \\ &\propto p(\theta) \prod_{t=1}^{N} \left[\theta^{x^t} (1-\theta)^{1-x^t} \right]. \end{aligned}$$

Estimating the Sex Ratio

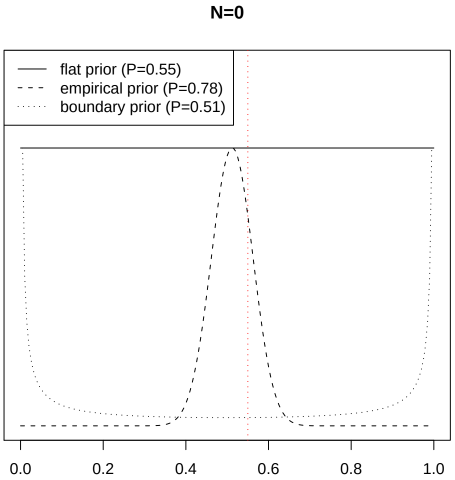
What is our degree of belief in the gender ratio, before seeing any data (prior probability density $p(\theta)$)?

- Very agnostic view: $p(\theta) = 1$ (flat prior).
- Something similar than elsewhere (empirical prior).
- Conspiracy theory prior: all newborns are almost all boys or all girls (boundary prior).



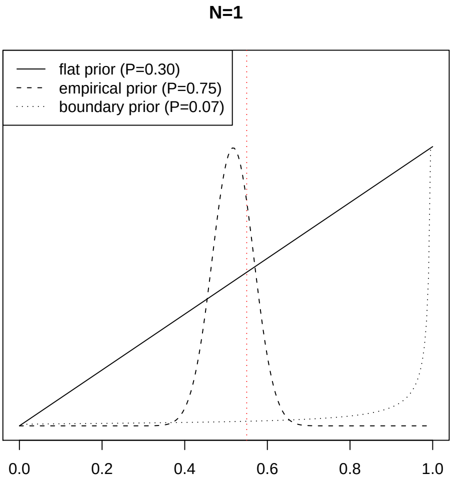
"True" $\theta = 0.55$ is shown by the red dotted line. The densities have been scaled to have a maximum of one.

Estimating the Sex Ratio
Posterior probability density



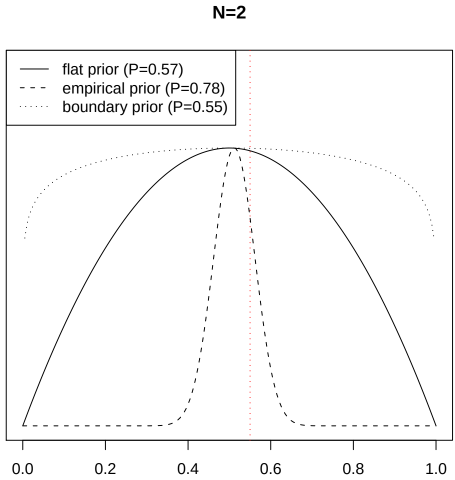
Estimating the Sex Ratio

Posterior probability density



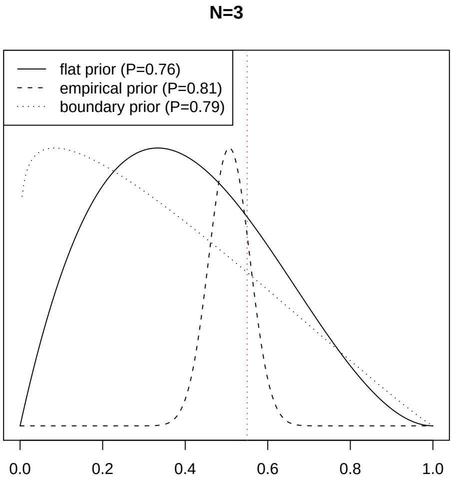
Estimating the Sex Ratio

Posterior probability density



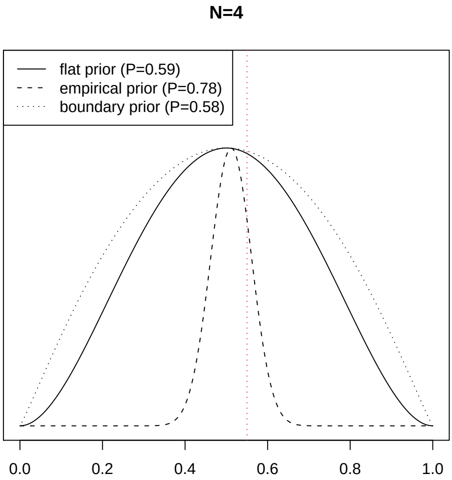
Estimating the Sex Ratio

Posterior probability density

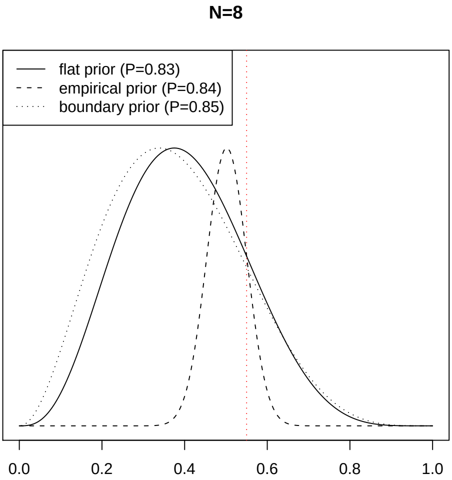


Estimating the Sex Ratio

Posterior probability density

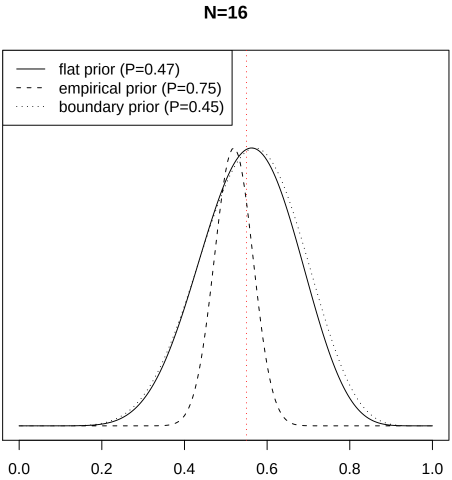


Estimating the Sex Ratio
Posterior probability density

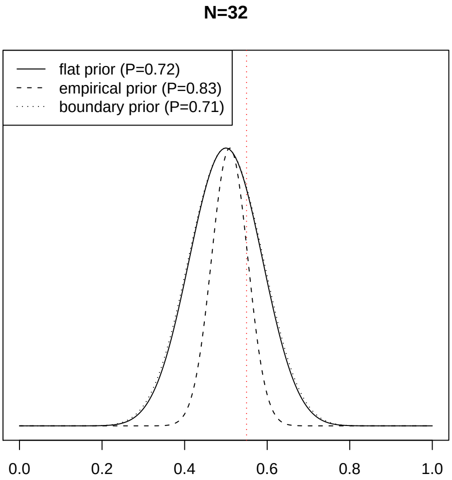


Estimating the Sex Ratio

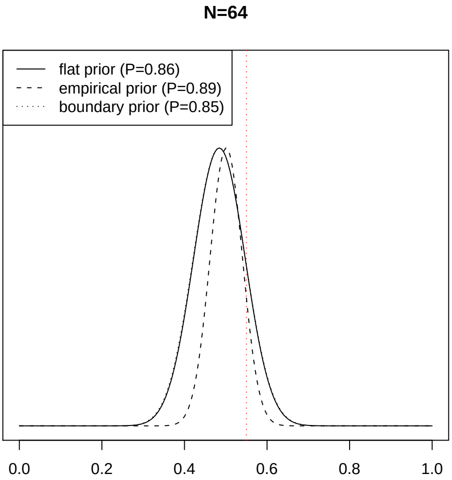
Posterior probability density



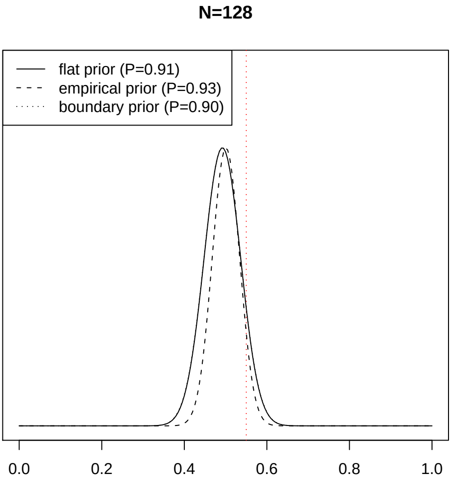
Estimating the Sex Ratio
Posterior probability density



Estimating the Sex Ratio
Posterior probability density

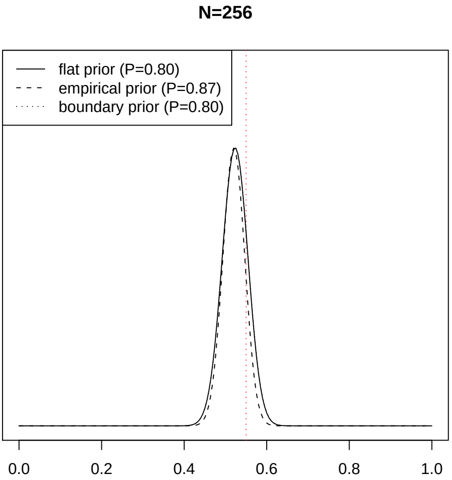


Estimating the Sex Ratio
Posterior probability density



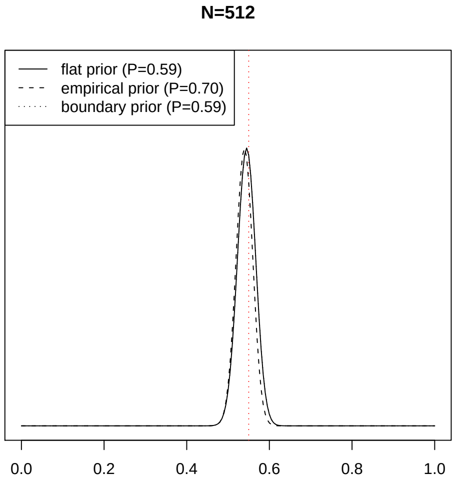
Estimating the Sex Ratio

Posterior probability density



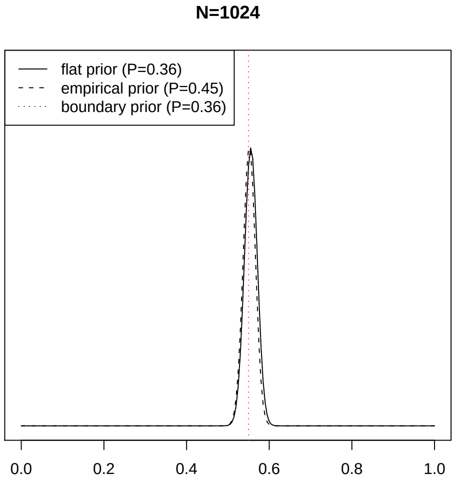
Estimating the Sex Ratio

Posterior probability density



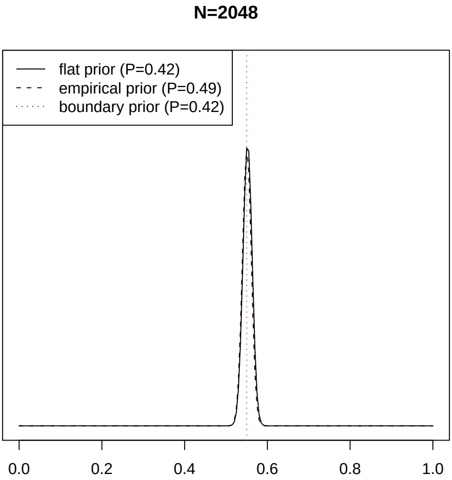
Estimating the Sex Ratio

Posterior probability density



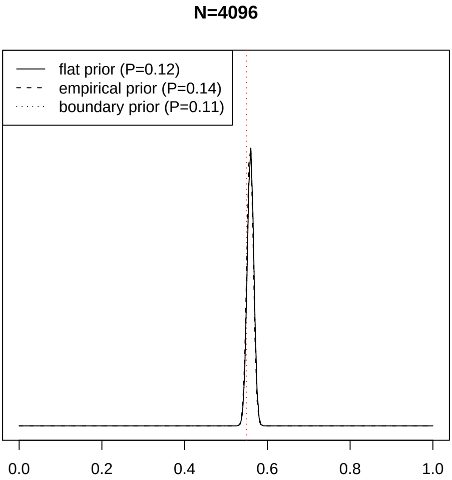
Estimating the Sex Ratio

Posterior probability density



Estimating the Sex Ratio

Posterior probability density



Observations

- With few data points the results are strongly dependent on the prior assumptions (inductive bias).
- As the number of data points grow, the results converge to the same answer.
- The conspiracy theory fades out quickly as we notice that there are both male and female babies.
- The only zero posterior probability is on hypothesis $\theta = 0$ and $\theta = 1$.
- It takes quite a lot observations to pin the result down to a reasonable accuracy.
- The posterior probability can be very small number. Therefore, we usually work with logs of probabilities.

Outline

Predictions from the Posterior

- The posterior represents our best knowledge.
- Predictor for new data point:

$$p(x \mid \mathcal{X}) = E_{p(\theta \mid \mathcal{X})}\left[p(x \mid \theta)\right] = \int d\theta p(x \mid \theta) p(\theta \mid \mathcal{X}).$$

- The calculation of the integral may be infeasible.
- Solution: estimate θ by $\hat{\theta}$ and use the predictor

$$p(x \mid \mathcal{X}) \approx p(x \mid \hat{\theta}).$$

Estimations from the Posterior

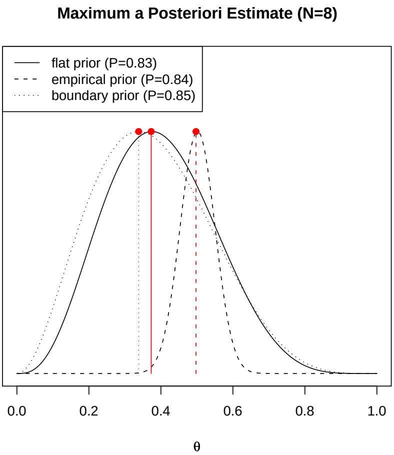
Definition (Maximum Likelihood Estimate)

$$\hat{\theta}_{ML} = \arg\max_{\theta} \log p(\mathcal{X} \mid \theta).$$

Definition (Maximum a Posteriori Estimate)

$$\hat{\theta}_{MAP} = \arg\max_{\theta} \log p(\theta \mid \mathcal{X}).$$

(With flat prior MAP Estimate reduces to the ML Estimate.)



Bernoulli Density

- Two states, $x \in \{0, 1\}$, one parameter $\theta \in [0, 1]$.

$$P(X = x \mid \theta) = \theta^x (1 - \theta)^{1-x} .$$

$$P(\mathcal{X} \mid \theta) = \prod_{t=1}^{N} \theta^{x^t} (1 - \theta)^{1-x^t} .$$

$$\mathcal{L} = \log P(\mathcal{X} \mid \theta) = \sum_t x^t \log \theta + \left(N - \sum_t x^t \right) \log (1 - \theta).$$

$$\frac{\partial \mathcal{L}}{\partial \theta} = 0 \Rightarrow \hat{\theta}_{ML} = \frac{1}{N} \sum_t x^t.$$

Multinomial Density

- K states, $x \in \{1, \ldots, K\}$, K real parameters $\theta_i \geq 0$ with constraint $\sum_{k=1}^{K} \theta_k = 1$.
- One observation is an integer k in $\{1, \ldots, K\}$ and it is represented by $x_i = \delta_{ik}$.

$$P(X = i \mid \theta) = \prod_{k=1}^{K} \theta_k^{x_k}.$$

$$P(\mathcal{X} \mid \theta) = \prod_{t=1}^{N} \prod_{k=1}^{K} \theta_k^{x_k^t}.$$

$$\mathcal{L} = \log P(\mathcal{X} \mid \theta) = \sum_{t=1}^{N} \sum_{k=1}^{K} x_k^t \log \theta_k.$$

$$\frac{\partial \mathcal{L}}{\partial \theta_k} = 0 \Rightarrow \hat{\theta}_{kML} = \frac{1}{N} \sum_t x_k^t.$$

Gaussian Density

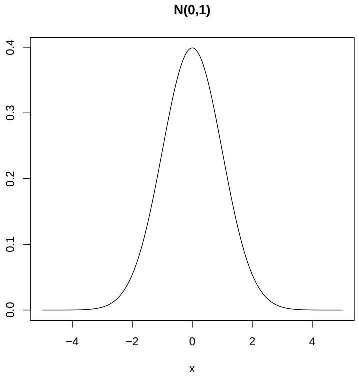
- A real number x is Gaussian (normal) distributed with mean μ and variance σ^2 or $x \sim N(\mu, \sigma^2)$ if its density function is

$$p(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

$$\mathcal{L} = \log P(\mathcal{X} \mid \mu, \sigma^2)$$

$$= -\frac{N}{2}\log(2\pi) - N\log\sigma - \frac{\sum_{t=1}^N (x^t-\mu)^2}{2\sigma^2}.$$

$$ML: \begin{cases} m = \frac{1}{N}\sum_{t=1}^N x^t \\ s^2 = \frac{1}{N}\sum_{t=1}^N (x^t - m)^2 \end{cases}$$



$$p(x \mid \mu = 0, \sigma^2 = 1)$$

Outline

Bias and Variance

- Setup: unknown parameter θ is estimated by $d(\mathcal{X})$ based on a sample $\mathcal{X}$.
- Example: estimate σ^2 by $d = s^2$.
- Bias: $b_\theta(d) = E[d] - \theta$.
- Variance: $E\left[(d - E[d])^2\right]$.
- Mean square error of the estimator $r(d, \theta)$:

$$
\begin{aligned}
r(d,\theta) &= E\left[(d-\theta)^2\right] \\
&= (E[d]-\theta)^2 + E\left[(d - E[d])^2\right] \\
&= \mathrm{Bias}^2 + \mathrm{Variance}.
\end{aligned}
$$

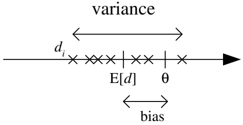


Figure 4.1 of Alpaydin (2004).

Bias and Variance

Unbiased estimator of variance

- Estimator is unbiased if $b_\theta(d) = 0$.
- Assume $\mathcal{X}$ is sampled from a Gaussian distribution.
- Estimate σ^2 by s^2: $s^2 = \frac{1}{N} \sum_t (x^t - m)^2$.
- We obtain:

$$E_{p(x|\mu,\sigma^2)}\left[s^2\right] = \frac{N-1}{N}\sigma^2.$$

- s^2 is not unbiased estimator, but $\frac{N}{N-1}s^2$ is:

$$\hat{\sigma}^2 = \frac{1}{N-1}\sum_{t=1}^{N}\left(x^t - m\right)^2.$$

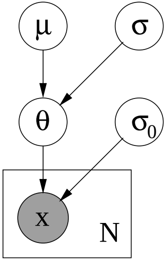
- s^2 is however asymptotically unbiased (that is, bias vanishes when $N \to \infty$).

Bayes' Estimator

- Bayes' estimator:
  $\hat{\theta}_{Bayes} = E_{p(\theta|\mathcal{X})}[\theta] = \int d\theta \theta p(\theta \mid \mathcal{X})$.
- Example: $x^t \sim N(\theta, \sigma_0^2)$, $t \in \{1, \ldots, N\}$, and $\theta \sim N(\mu, \sigma^2)$, where μ, σ^2 and σ_0^2 are known constants. Task: estimate θ.

$$p(\mathcal{X} \mid \theta) = \frac{1}{(2\pi\sigma_0^2)^{N/2}} \exp\left(-\frac{\sum_t (x^t - \theta)^2}{2\sigma_0^2}\right),$$

$$p(\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\theta - \mu)^2}{2\sigma^2}\right).$$



- It can be shown that $p(\theta \mid \mathcal{X})$ is Gaussian distributed with

$$\hat{\theta}_{Bayes} = E_{p(\theta|\mathcal{X})}[\theta] = \frac{N/\sigma_0^2}{N/\sigma_0^2 + 1/\sigma^2} m + \frac{1/\sigma^2}{N/\sigma_0^2 + 1/\sigma^2} \mu.$$

Outline

About Estimators

- Point estimates collapse information contained in the posterior distribution into one point.
- Advantages of point estimates:
    - Computations are easier: no need to do the integral.
    - Point estimate may be more interpretable.
    - Point estimates may be good enough. (If the model is approximate anyway it may make no sense to compute the integral exactly.)
- Alternative to point estimates: do the integral analytically or using approximate methods (MCMC, variational methods etc.).
- One should always use test set to validate the results. The best estimate is the one performing best in the validation/test set.

Conclusion

- Next lecture: More about Model Selection (Alpaydin (2004) Ch 4)
- Problem session on 5 October.