PMR Learning as Inference

Probabilistic Modelling and Reasoning

Amos Storkey

School of Informatics, University of Edinburgh

Outline

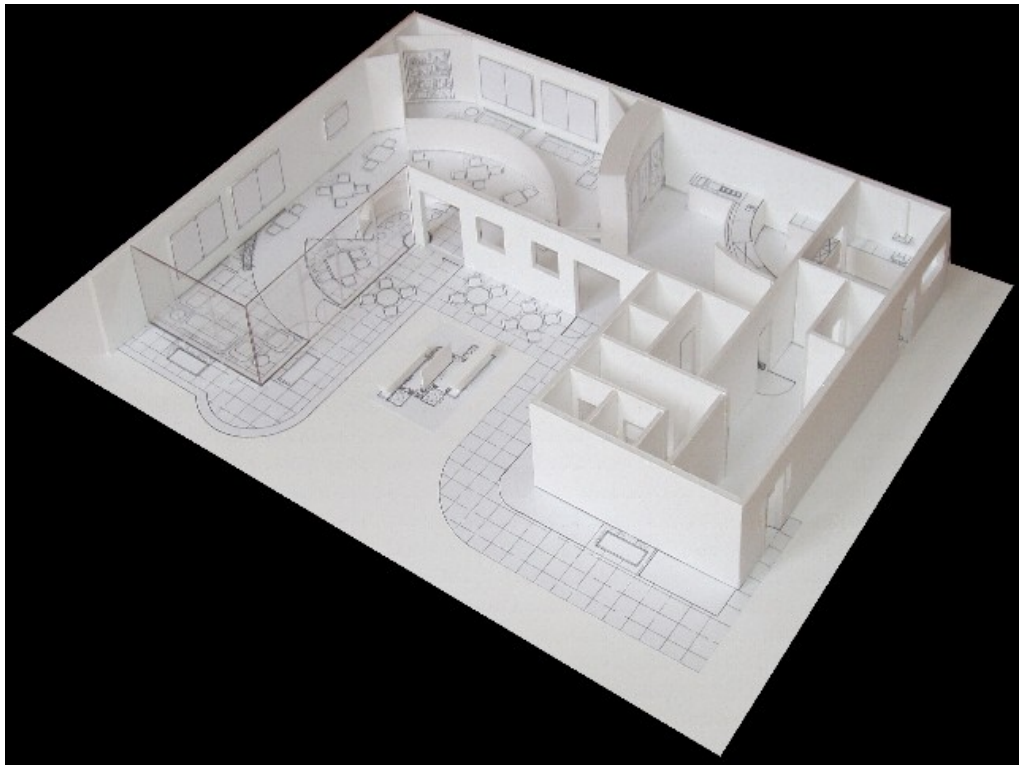
Modelling

Probabilistic Modelling is about building models and using them.

What do we mean by modelling?

Modelling

Modelling



A Generative Model

- Building an idealisation to capture the essential elements of an item.
- Think of a model as a model for future data generation
- Given a model (a distribution) we can sample from that distribution to get artificial data.
- Need to specify enough to do this generation.
- Often make IID (Independent and Identically Distributed) Assumption
- Models are not truth. They try to capture our uncertainties.

The Inverse Problem

- We built a generative model, or a set of generative models on the basis of what we know (prior).
- Can generate artificial data.
- BUT what if we want to *learn* a good distribution for data that we then see? How is goodness measured?

Explaining Data

A particular distribution explains the data better if the data is more probable under that distribution.

- The likelihood approach

Likelihood

- $P(\mathcal{D}|\mathcal{M})$. The probability of the data $\mathcal{D}$ given a distribution (or model) $\mathcal{M}$. This is called the likelihood of the model.
- This is

$$P(\mathcal{D}|\mathcal{M}) = \prod_{n=1}^{N} P(x^n|\mathcal{M})$$

  i.e. the product of the probabilities of generating each data point individually.
- This is a result of the independence assumption (indep $\rightarrow$ product of probabilities by definition).
- Try different $\mathcal{M}$ (different distributions). Pick the $\mathcal{M}$ with the highest likelihood $\rightarrow$ Maximum Likelihood Approach.

Bernoulli model

Example

Data: 1 0 0 1 0 1 0 1 0 0 0 0 0 1 0 1 1 1 0 1.

- Continuous range of hypotheses: $\mathcal{M} = p$ - Generated from a Boolean distribution with $P(1|p) = p$.
- Likelihood of data. Let c=number of ones:

$$\prod_{n=1}^{N} P(x^n|p) = p^c(1-p)^{20-c}$$

- Maximum likelihood hypothesis? Differentiate w.r.t. p to find maximum
- In fact usually easier to differentiate $\log P(\mathcal{D}|\mathcal{M})$: $\log$ is monotonic. So argmax $\log(f(x))$ = argmax $f(x)$.

Bernoulli model

Example

Data: 1 0 0 1 0 1 0 1 0 0 0 0 0 1 0 1 1 1 0 1.

- Likelihood of data. Let c=number of ones:

$$\log \prod_{n=1}^{N} P(x^n|p) = c \log p + (20 - c) \log(1 - p)$$

- Set $d/dp \log P(\mathcal{D}|\mathcal{M}) = c/p - (20 - c)/(1 - p)$ to zero to find maximum.
- So $c(1 - p) - (20 - c)p = 0$. This gives $p = c/20$. Maximum likelihood is unsurprising.

Distributions over Parameters

Parameter bias

- Although we are uncertain about the parameter values, often some are more probable than others.
- Uncertainty $\rightarrow$ probability: put *prior* (distribution) on parameters.
- Compute max *posterior* instead of max likelihood
- Bayes Rule:

$$\text{Posterior} \rightarrow P(p|\mathcal{D}) = \frac{P(\mathcal{D}|p)P(p)}{P(\mathcal{D})} \leftarrow \text{Prior}$$

- $P(\mathcal{D}) = \int dp\ P(p|\mathcal{D})P(p)$ does not depend on p.
- argmax $P(\mathcal{D}|p)P(p)$
- argmax $(\log P(\mathcal{D}|p) + \log P(p)) \leftarrow$ penalty term

Maximum Posterior

Example

1 0 0 1 0 1 0 1 0 0 0 0 0 1 0 1 1 1 0 1. $P(p) \propto p(1-p)$

- Let c=number of ones (9). Then $\log P(p|\mathcal{D}) =$

$$c \log p + (20 - c) \log(1 - p) + \log p + \log(1 - p) + \text{const}$$

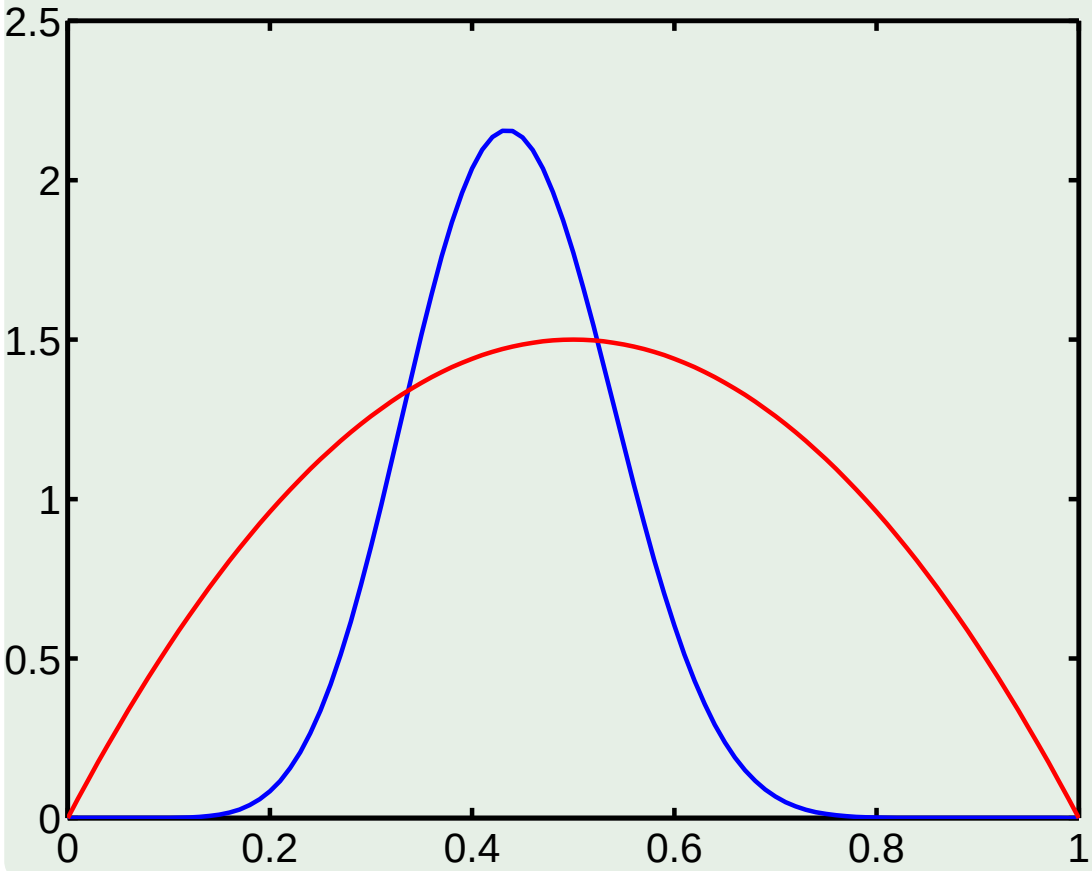
- Set $d/dp \log P(\mathcal{D}|\mathcal{M}) = (c+1)/p - (20 - c + 1)/(1 - p)$ to zero to find maximum.
- So $(c + 1)(1 - p) - (20 - c + 1)p = 0$. This gives $p = (c + 1)/22 = 9/22 \approx 0.41$.
- With this prior, max. posterior prefers p closer to $1/2$.

Uncertainty of Parameters

- Maximizing the posterior gets us one value for the parameter.
- Is it right?
- No it is an estimate. But how good an estimate? There is some uncertainty. How much?
- Uncertainty $\rightarrow$ probability.
- Find posterior *distribution* over parameters, not just maximum.

Posterior Distribution

Example



Prior in Red,
Posterior in Blue

Inference and Marginalisation

Example

1 0 0 1 0 1 0 1 0 0 0 0 0 1 0 1 1 1 0 1. $P(p) \propto p(1-p)$

- What is probability of next item x^* being 1? Predict.
- Could take maximum posterior parameter and compute probability of next item? (≈ 0.41)
- But: lots of possible posterior parameters. Some more possible than others.
- Instead marginalise:

$$\int dp \; P(x_* = 1|p)P(p|\mathcal{D})$$

- This gives approximately 0.46.

Test

- We considered choosing between model, where each model defined a precise distribution.
- But what if each model defines a whole *type* of distribution.
- We might not know the precise *parameters* of the distribution.
- Compute the *evidence* or *marginal likelihood*:
- Marginalise out the unknown parameters to get likelihood of model.

Learning as inference

- Its just as if the parameters were nodes in our graphical model.
- In fact that is exactly what they are.
- Latent variables - intrinsic - separate variables for each data item.
- Parameters - extrinsic - shared across all data items.

Summary of Bayesian Computation

- Define prior model $P(\mathcal{D})$, usually by using

$$P(\mathcal{D}) = \int d\boldsymbol{\theta}\ P(\mathcal{D}|\boldsymbol{\theta})P(\boldsymbol{\theta})$$

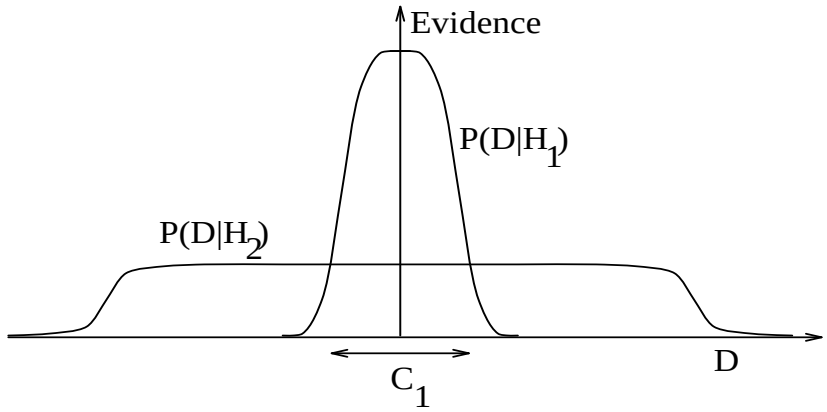
  and defining:
  - The likelihood $P(\mathcal{D}|\boldsymbol{\theta})$ with parameters $\boldsymbol{\theta}$.
  - The *prior distribution* (over parameters) $P(\boldsymbol{\theta}|\boldsymbol{\alpha})$ which might also be parameterized by hyper-parameters $\boldsymbol{\alpha}$.
- Conditioning on data to get the *posterior distribution* over parameters $P(\boldsymbol{\theta}|\mathcal{D})$.
- Using the posterior distribution for prediction (inference)

$$P(\mathbf{x}^*|\mathcal{D}) = \int d\boldsymbol{\theta}\ P(\mathbf{x}^*|\boldsymbol{\theta})P(\boldsymbol{\theta}|\mathcal{D})$$

Why not maximize?

We have described learning as an inference procedure. But why not maximize.

- Why be Bayesian? Why not compute best parameters and compare?
- More parameters=better fit to data. ML: bigger is better.
- But might be overfitting: only these parameters work. Many others don't.



- Prefer models that are unlikely to 'accidentally' explain the data.
- That said, maximum posterior parameters are often good

Recap

- For Bernoulli likelihood with Beta prior, could do Bayesian computation analytically.
- For Binomial likelihood and Beta prior, could do Bayesian computation analytically.
- For Multinomial likelihood and Dirichlet prior, could do Bayesian computation analytically.
- Question: are there other distributions for which we can do analytical Bayesian computations?
- Is this a good thing? Discuss.

Analytical methods

- Yes: conjugate exponential models.
- Good thing: easy to do the sums.
- Bad thing: prior distribution should match beliefs. Does a Beta distribution match your beliefs? Is it good enough?
- Certainly not always.

The exponential family

- Any distribution over some $\mathbf{x}$ that can be written as

$$P(\mathbf{x}|\boldsymbol{\eta}) = h(\mathbf{x})g(\boldsymbol{\eta})\exp\left(\boldsymbol{\eta}^T\mathbf{u}(\mathbf{x})\right)$$

  with h and g known, is in the *exponential family* of distributions.
- Many of the distributions we have seen are in the exponential family. A notable exception is the t-distribution.
- The $\boldsymbol{\eta}$ are called the *natural parameters* of the distribution.

Wait - I didn't get that!

$$P(\mathbf{x}|\boldsymbol{\eta}) = h(\mathbf{x})g(\boldsymbol{\eta})\exp\left(\boldsymbol{\eta}^T\mathbf{u}(\mathbf{x})\right)$$

- More simply....
- Any distribution that can be written such that the interaction term (between parameters and variables) is log linear in the parameters is in the *exponential family*.
- i.e.

$$\log P(\mathbf{x}|\boldsymbol{\eta}) = \sum_i \eta_i u_i(\mathbf{x}) + (\text{other stuff that only contains } \mathbf{x} \text{ or } \boldsymbol{\eta})$$

- A distribution may usually be parameterized in a way that is different from the exponential family form.
- So sometimes useful to convert to exponential family representation and find the 'natural' parameters.

The exponential family

- Multinomial Distribution

$$P(\mathbf{x}|\{\log p_k\}) \propto \exp\left(\sum_k x_k \log p_k\right)$$

The Gaussian Distribution

- Need to intro the Gaussian first.

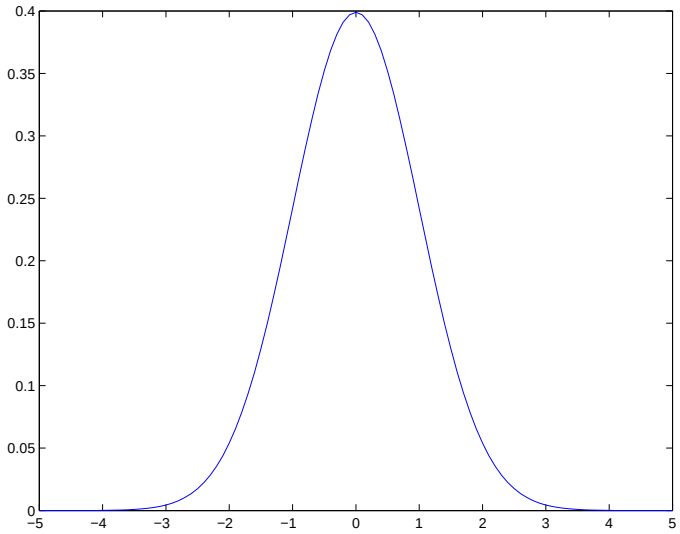
Definition

- The one dimensional Gaussian distribution is given by

$$P(x|\mu,\sigma^2) = N(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp -\frac{(x-\mu)^2}{2\sigma^2}$$

- μ is the *mean* of the Gaussian and σ^2 is the *variance*.
- If $\mu = 0$ and $\sigma^2 = 1$ then $N(x;\mu,\sigma^2)$ is called a *standard* Gaussian.

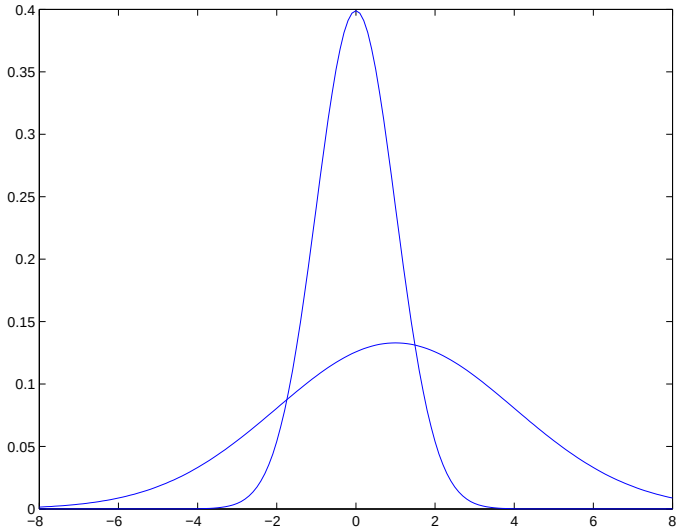
Plot



- This is a standard one dimensional Gaussian distribution.
- All Gaussians have the same shape subject to scaling and displacement.
- If x is distributed $N(x; \mu, \sigma^2)$, then $y = (x - \mu)/\sigma$ is distributed $N(y; 0, 1)$.

Normalisation

- Remember all distributions must integrate to one. The $\sqrt{2\pi\sigma^2}$ is called a normalisation constant - it ensures this is the case.
- Hence tighter Gaussians have higher peaks:



Central Limit Theorems

- X_i mean 0, variance Σ, not necessarily Gaussian.
- X_i subject to various conditions (e.g. IID, light tails).

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{N} X_i \sim N(0, \Sigma)$$

asymptotically as $N \to \infty$.

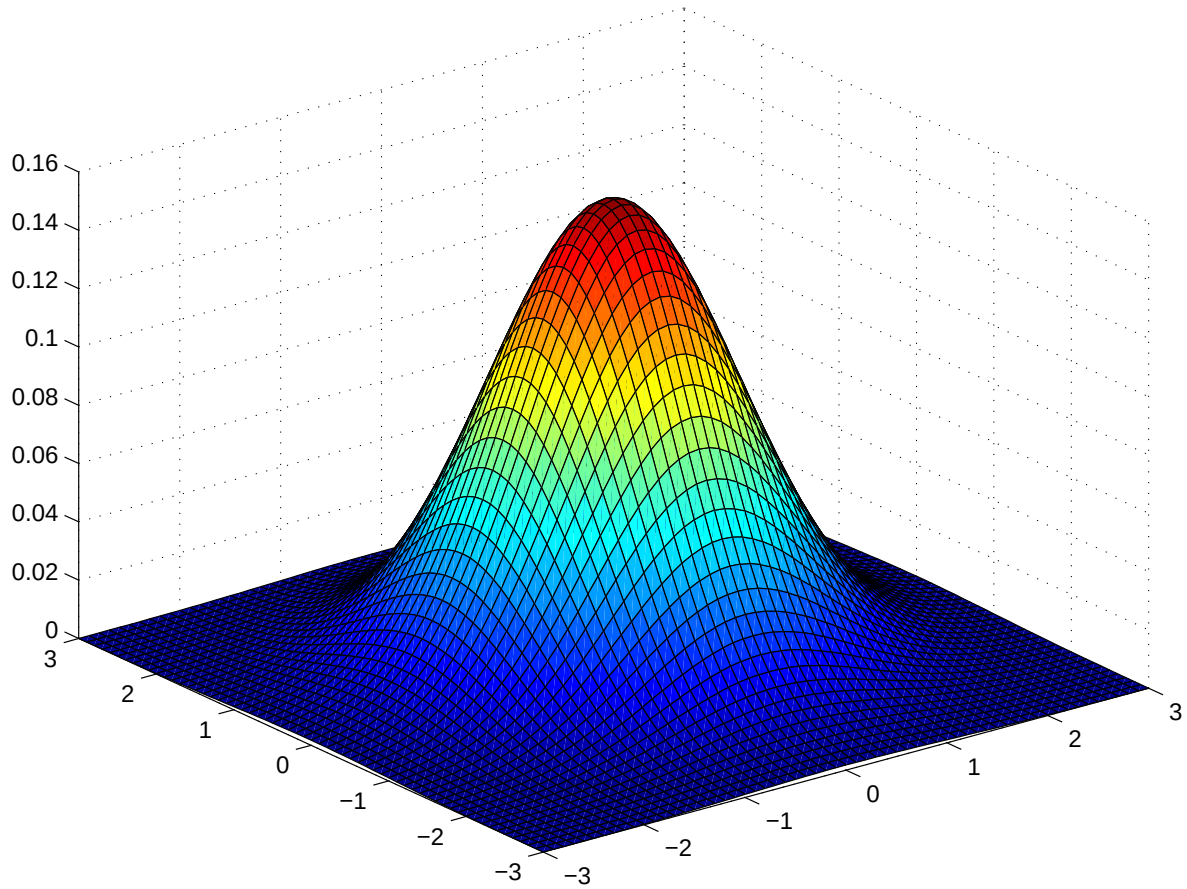
Multivariate Gaussian

- The vector $\mathbf{x}$ is multivariate Gaussian if for mean μ and covariance matrix Σ, it is distributed according to

$$P(\mathbf{x}|\mu,\Sigma) = \frac{1}{|(2\pi)\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)\right)$$

- The univariate Gaussian is a special case of this.
- Σ is called a covariance matrix. It says how much attributes co-vary. More later.

Multivariate Gaussian: Picture



The exponential family

- Gaussian Distribution

$$P(\mathbf{x}|\boldsymbol{\eta}) \propto \exp\left(\sum_k \eta_k x_k - \frac{1}{2}\sum_{ij} \Sigma_{ij}^{-1} x_i x_j)\right)$$

- $\eta = \Sigma^{-1}\mu$.

Pause

Conjugate exponential models

- If the prior takes the same functional form as the posterior for a given likelihood, a prior is said to be *conjugate* for that likelihood.
- There is a conjugate prior for any exponential family distribution.
- If the prior and likelihood are conjugate and exponential, then the the model is said to be *conjugate exponential*
- In conjugate exponential models, the Bayesian integrals can be done analytically.

Conjugacy

- In high dimensional spaces it is hard to accurately estimate the parameters using maximum likelihood. Can utilise Bayesian methods.
- Conjugate distribution for the Gaussian with mean parameter is another Gaussian.
- Conjugate distribution for the Gaussian with precision (inverse variance) parameter is the Gamma distribution.
- Conjugate distribution for the Gaussian with precision matrix (inverse covariance is the Wishart distirbution.
- Conjugate distribution for the Gaussian with both mean and precision matrix is the Gaussian-Wishart distribution.
- Wishart distribution is distribution over matrices!

Conjugacy

- Remember - for conjugate distribution posterior is of the same form.
- So given the data, we just need to update the hyperparameters of the prior distribution to get the posterior.

Example

- Gaussian $N(\mu, \Lambda^{-1})$. Fixed precision Λ, but μ distributed $N(\mu_0, \Lambda_0^{-1})$
- Posterior mean $(\Lambda_0 + n\Lambda)^{-1}(\Lambda_0\mu_0 + n\Lambda\bar{x})$
- Posterior precision $(\Lambda_0 + n\Lambda)$.

Conjugacy: Evidence

Example

- Gaussian likelihood $N(\mathbf{x}; \boldsymbol{\mu}, \Sigma)$, Gaussian prior $N(\boldsymbol{\mu}; \boldsymbol{\mu}_0, \Sigma_0)$. Simple case: $\boldsymbol{\mu}_0 = 0$, Σ known.
- Marginal likelihood (Evidence)? We know Marginal Likelihood is Gaussian. So using $\mathbf{x} = \boldsymbol{\mu} + \boldsymbol{\epsilon}$, $\boldsymbol{\epsilon}$ mean 0, covariance Σ, compute mean $\mathbf{m}$ and covariance C of marginal likelihood

$$\mathbf{m} = \langle \mathbf{x} \rangle = \langle \boldsymbol{\mu} \rangle + 0 = 0$$

$$C = \langle \mathbf{x}\mathbf{x}^T \rangle = \langle \boldsymbol{\mu}\boldsymbol{\mu}^T \rangle + \langle \boldsymbol{\epsilon}\boldsymbol{\epsilon}^T \rangle + 0 = \Sigma + \Sigma_0$$

Give me more...

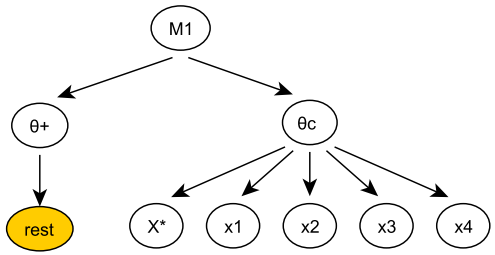
- Red Orange Yellow Aquamarine
- Haggis Mountains Loch Celtic Castle
- Trees, Forests, Pruning, Parent, Machine Learning, Bayesian.
- Google Sets

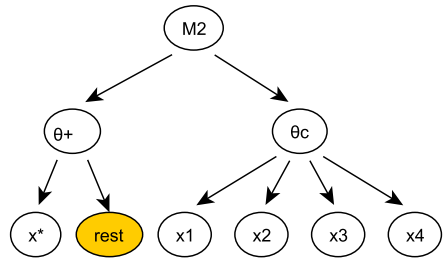
Different features

- Have a large database of objects, each described by $\mathcal{D}^+$ (e.g. Web)
- Have a small number of examples from the dataset, each with various (binary) features, which we collect into $\mathcal{D}_c$.
- Want to pick things from $\mathcal{D}^+$ that 'belong to the same set' as those in $\mathcal{D}_c$
- How should we do it?

Model

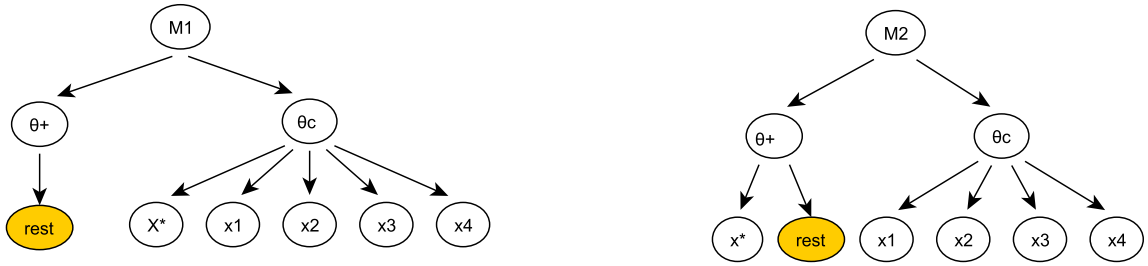
- Data consists of $\mathcal{D}_c$ and query point $\mathbf{x}^*$. Denote by $\mathcal{D}$.
- Two models: $\mathcal{M}_1$: $\mathcal{D}$ all from same subset C, or $\mathcal{M}_2$: $\mathcal{D}_c$ from the same subset C, but $\mathbf{x}$ from the general distribution over all data $\mathcal{D}^+$





- Parameter vector is vector of (Boolean) probabilities, one for each feature.
- $\mathcal{D}^+$ is vast, and so presume maximum likelihood estimate good enough for $\mathcal{M}1$: have vector θ^+ for this.

Score



- Parameter vector θ_c for subset C is not known. So put a conjugate prior on the parameters: a Beta distribution for each component i of the feature vector, with hyper-parameters a_i and b_i.
- Compute $P(\mathcal{D}|\mathcal{M}_1)/P(\mathcal{D}|\mathcal{M}_2)$ (called the Bayes Factor).
- The larger this ratio is, the more this favours $\mathbf{x}^*$ being included in the set.
- Bayesian Model Comparison: parameters integrated out:

$$P(\mathcal{D}|\mathcal{M}_2) = \int P(\mathcal{D}|\theta)P(\theta|\alpha)d\theta$$