iPMR

Example

Example

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

- *p* denotes the probability of a 1 turning up. What is the probability that the next item x^* is 1?
- *p* takes one value for all data items. But we don't know what that is! Use a distribution to represent and compute with the uncertainty.

$$\int dp \; P(x_* = 1|p)P(p|D) \text{ where } P(p|D) = \frac{P(D|p)P(p)}{P(D)}.$$

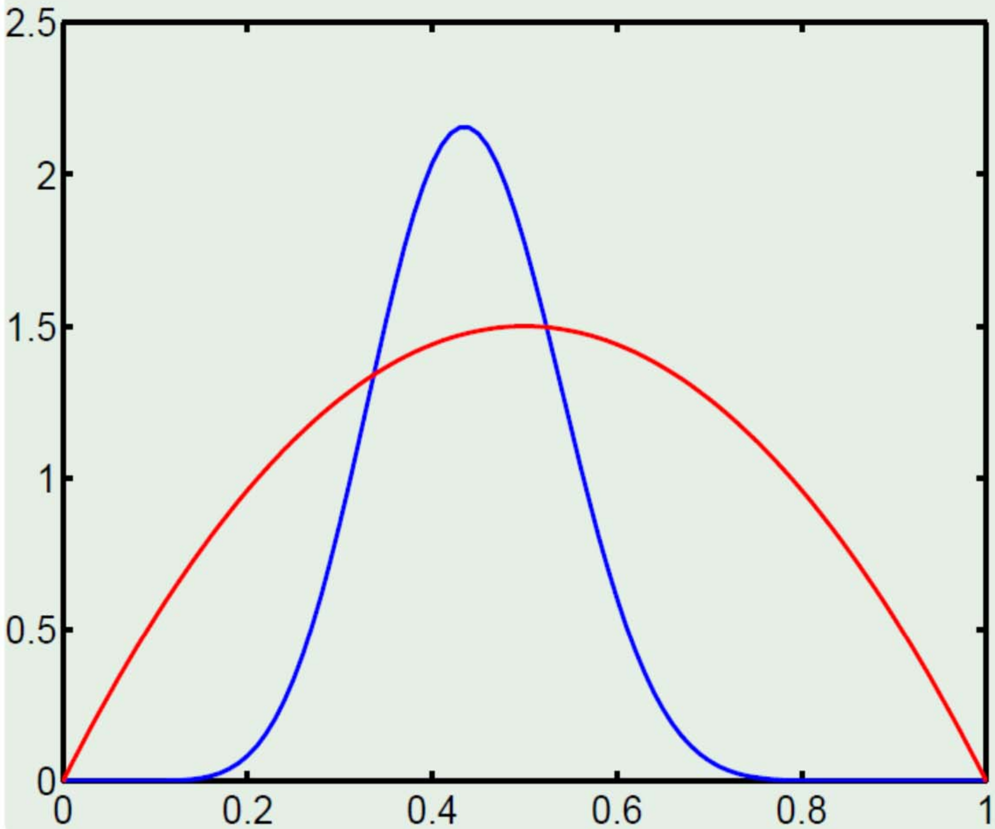
$$\text{Now } P(D|p) = \prod_n P(x^n|p) = p^{N_1}(1-p)^{N_0} = p^9(1-p)^{11}$$

$$\text{so } P(p|D) = \frac{23!}{10!12!}p^{10}(1-p)^{12}$$

meaning $P(x_* = 1) = \int dp \; P(x_* = 1|p)P(p|D)$. Computing this gives $11/24$. (Hint: look up Beta distribution).

Note $N_1 = 9$ is the number of ones, and $N_0 = 11$ is the number of zeros.

Example



Prior in Red,
Posterior in Blue

Summary of Bayesian Methods

- 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)

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

Question

- For Bernoulli likelihood with Beta prior, can do Bayesian computation analytically.
- For Binomial likelihood and Beta prior, can do Bayesian computation analytically.
- For Multinomial likelihood and Dirichlet prior, can do Bayesian computation analytically.
- Are there other situations this holds?

Conjugacy

- 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?

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 η are called the *natural parameters* of the distribution.

Exponential Family

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

- 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.

Exponential Family

- E.g. Multivariate distribution x

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

The Gaussian

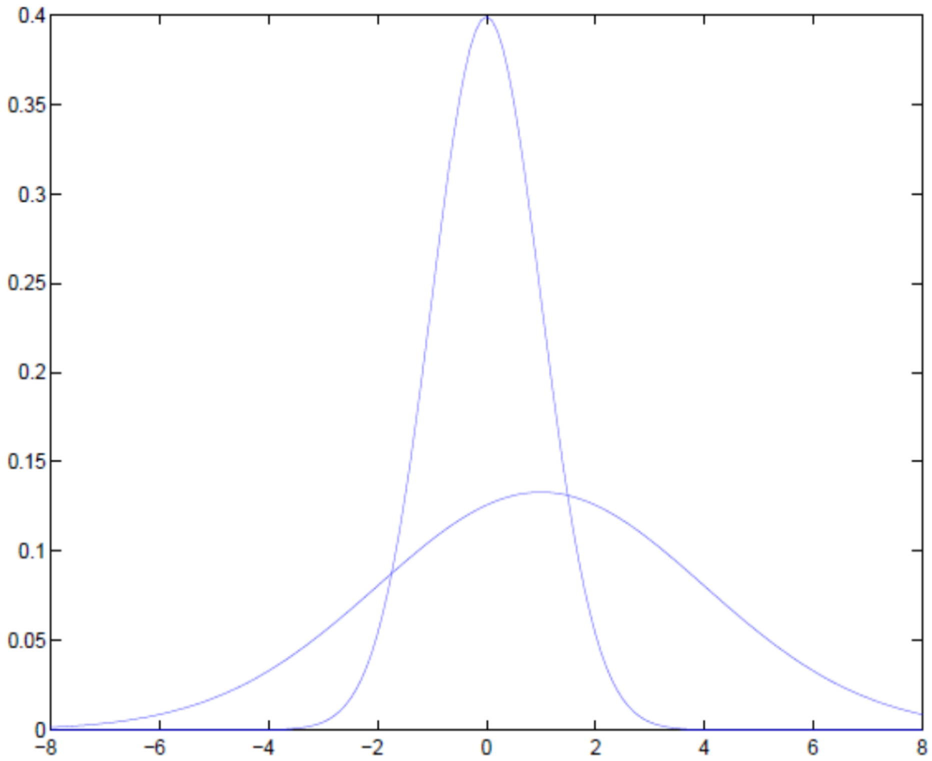
- The one dimensional Gaussian distribution is given by

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

- 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.

Gaussians

- Remember the normalisation (wider versus taller).
- Remember we can remap to a standard normal: $y = (x - \mu)/\sigma$



Multivariate Gaussian

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

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

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

Gaussian is in Exponential Family

- Gaussian Distribution

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

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

Conjugate Exponential

- 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 model is said to be conjugate exponential
- In conjugate exponential models, the Bayesian integrals can be done analytically.
- Update rules for conjugate distributions.

Example

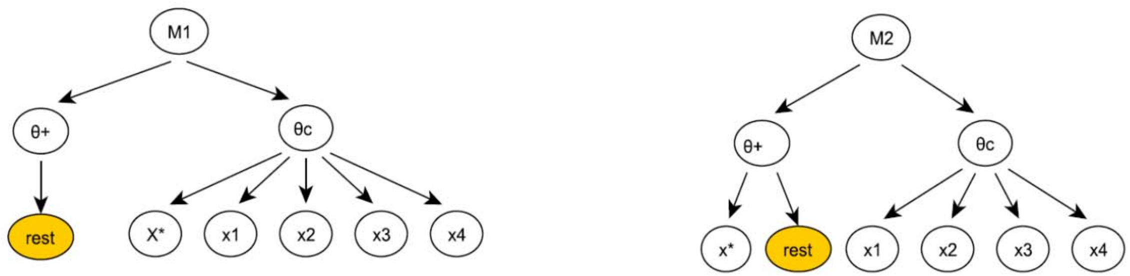
- What other items are in the set:
  - Red Orange Yellow Aquamarine
  - Haggis Mountains Loch Celtic Castle
  - Trees, Forests, Pruning, Parent, Machine Learning, Bayesian.

Bayesian Sets

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

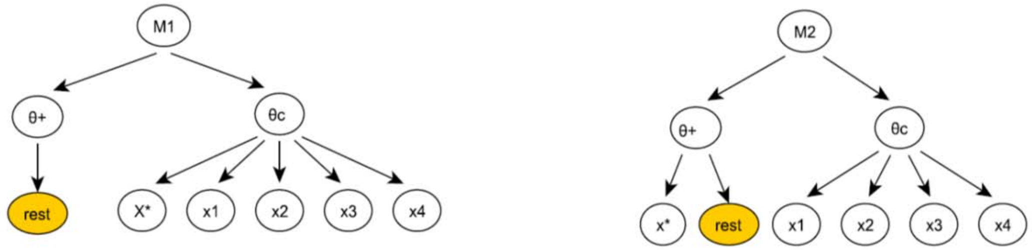
Bayesian Sets

- 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.

Bayesian Sets



- 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:

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

Our Journey

Graphical Models | Learning Probabilistic Models

- Introduction to Learning
- Learning exponential family models and Bayesian Set example.
- Next: Approximate Learning and Maximum Likelihood