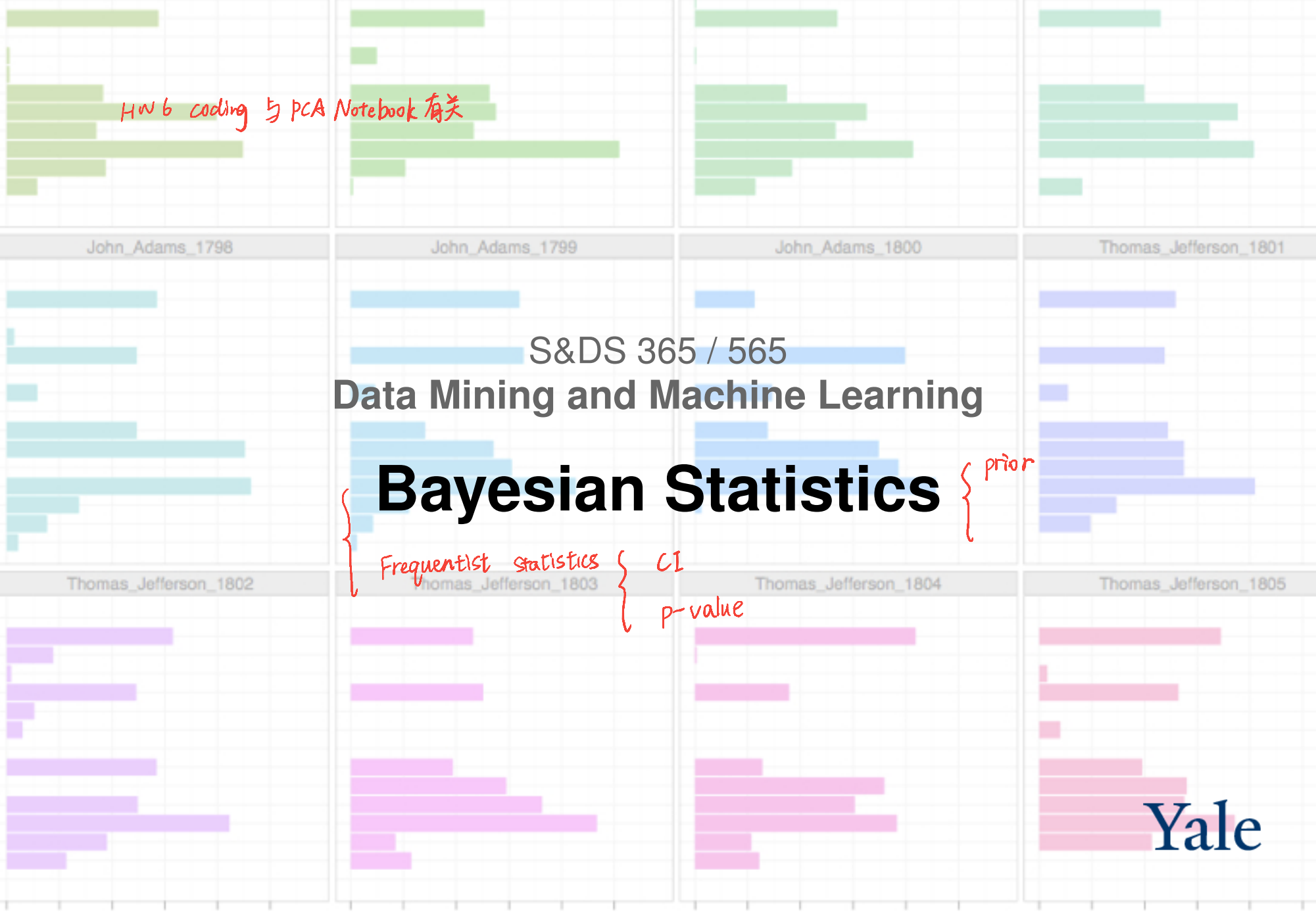


S&DS 365 / 565

Data Mining and Machine Learning

Bayesian Statistics

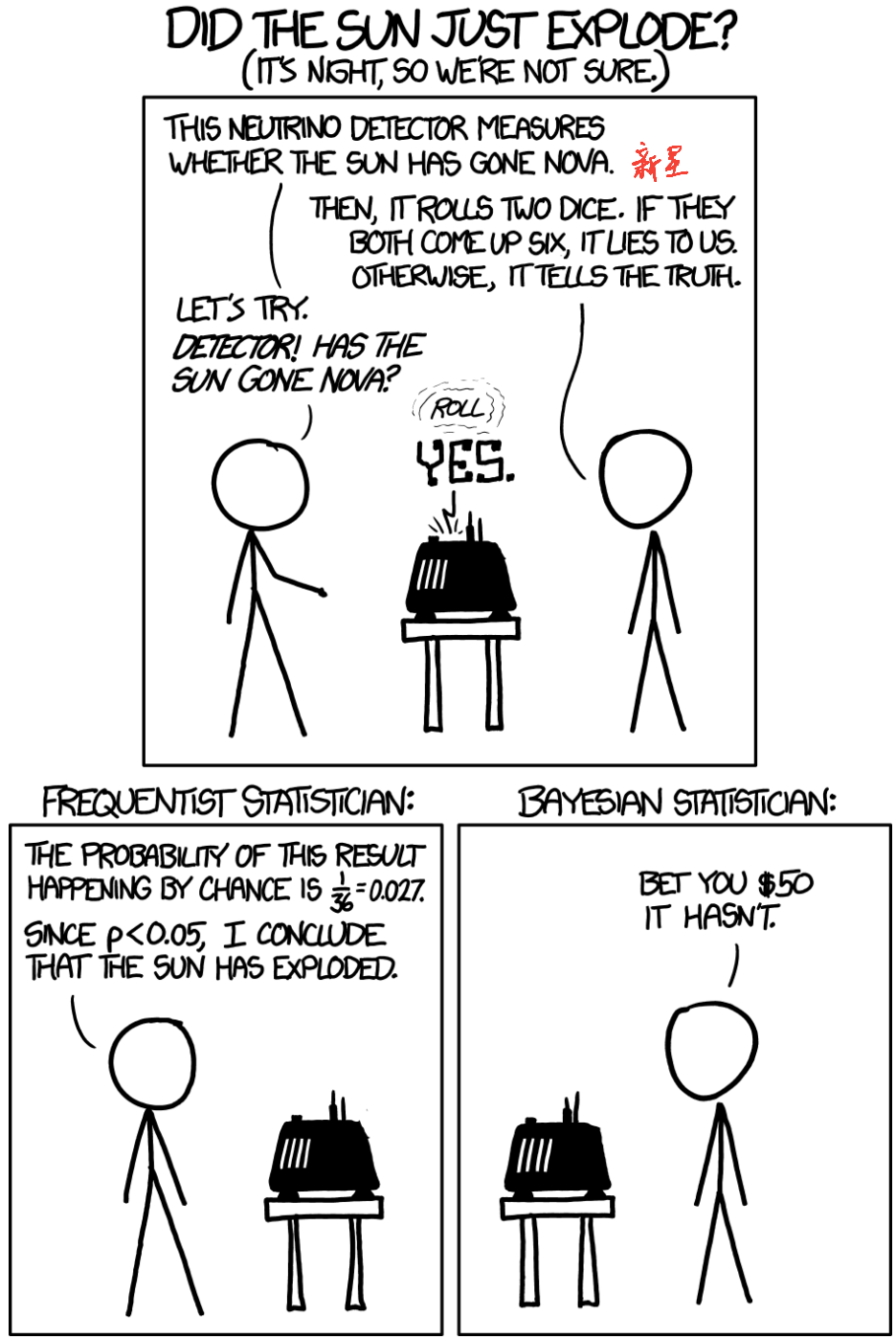
prior

Frequentist statistics

CI

p-value

Yale

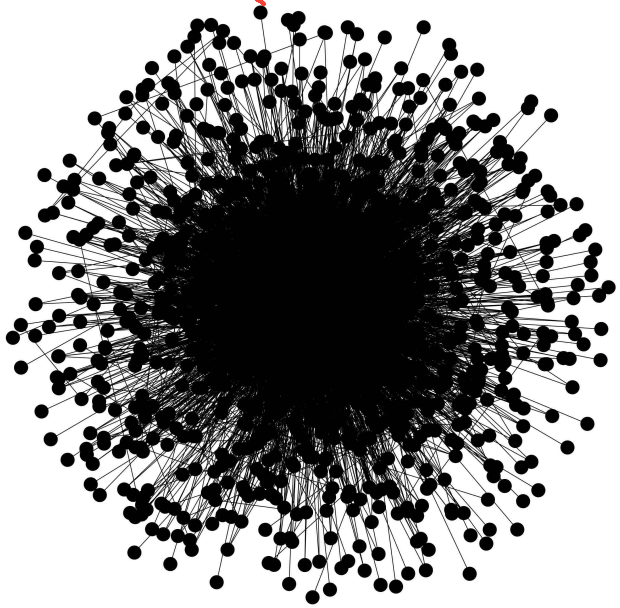

DID THE SUN JUST EXPLODE?
(IT'S NIGHT, SO WE'RE NOT SURE.)
THIS NEUTRINO DETECTOR MEASURES WHETHER THE SUN HAS GONE NOVA.
新星
THEN, IT ROLLS TWO DICE. IF THEY BOTH COME UP SIX, IT LIES TO US. OTHERWISE, IT TELLS THE TRUTH.
LET'S TRY.
DETECTOR! HAS THE SUN GONE NOVA?
ROLL
YES.
FREQUENTIST STATISTICIAN:
THE PROBABILITY OF THIS RESULT HAPPENING BY CHANCE IS 1/36 = 0.027.
SINCE p<0.05, I CONCLUDE THAT THE SUN HAS EXPLODED.
BAYESIAN STATISTICIAN:
BET YOU $50 IT HASN'T.

Idea: Community Detection

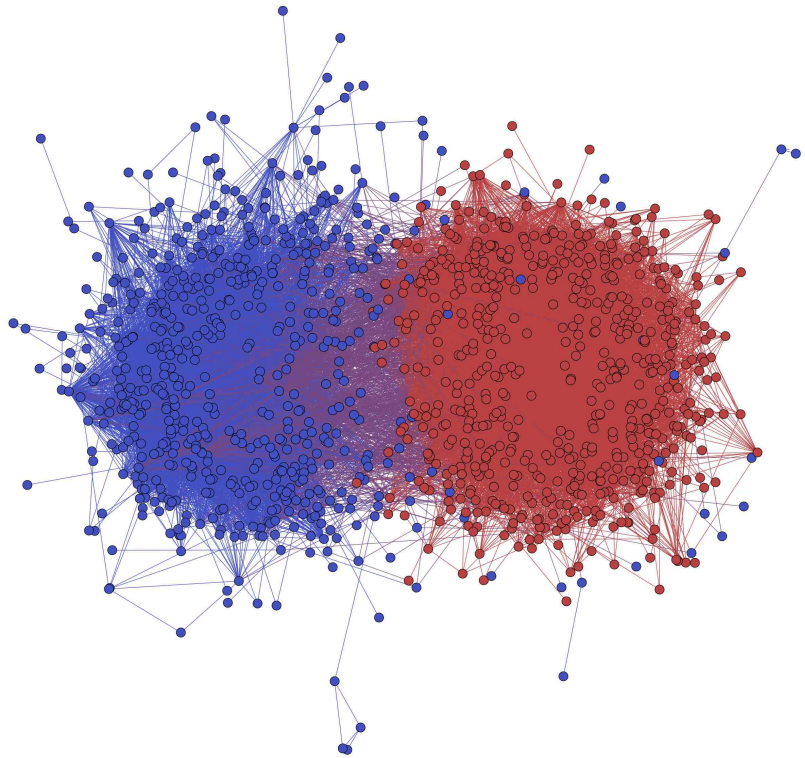
used to identify latent layer of data

graphs

blogs are connected with each other



clustering { PCA
Baysian



Idea: Finding structure in music

PCA gave us a way to find structure in the data.

This structure isn't necessarily a latent structure (some people would say it is)

Often latent structure is modeled as random.

For instance generating music: first randomly pick a style, then an artist, then generate the music

genere

In this case observed data is the music. Latent (potentially unknown variables) include the artist and style (e.g. shazam figuring out artist)

歌曲识别软件

Spotify uses machine learning to find relationships between music (identifying properties of the music versus hand creating features)

Pandora

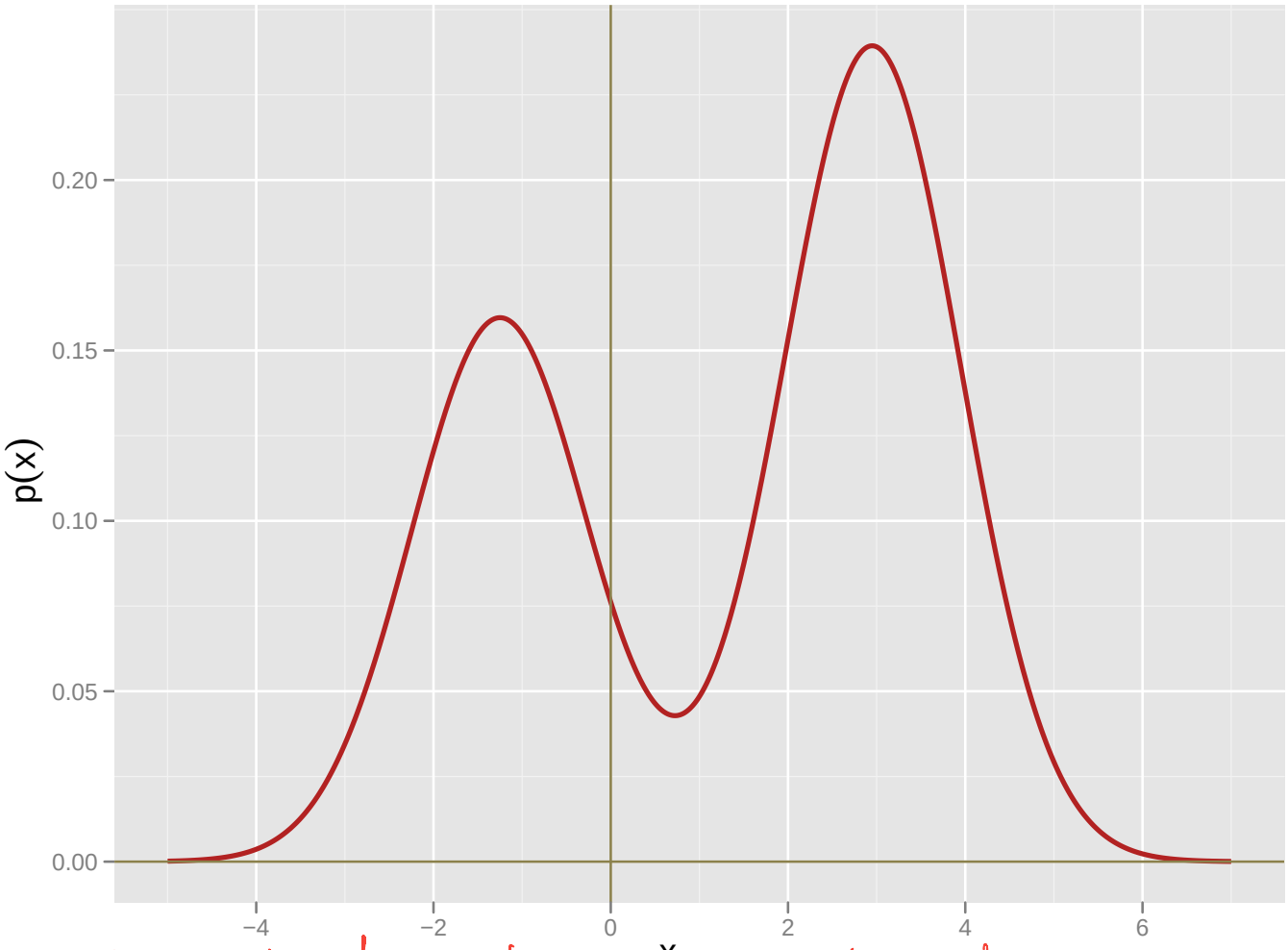
Mixtures

- Key technique: Mixture models
- Mixtures have latent variables
- Flexible tool
- S1mple and difficult at the same time

Gaussian Mixture

$X_1 \sim N(-1.25, 1)$ $X_2 \sim N(2.95, 1)$ π_i: probability of observing X_i

$\pi_1 = \frac{2}{5}$ $\pi_2 = \frac{3}{5}$ X_i: mass of a butterfly species i



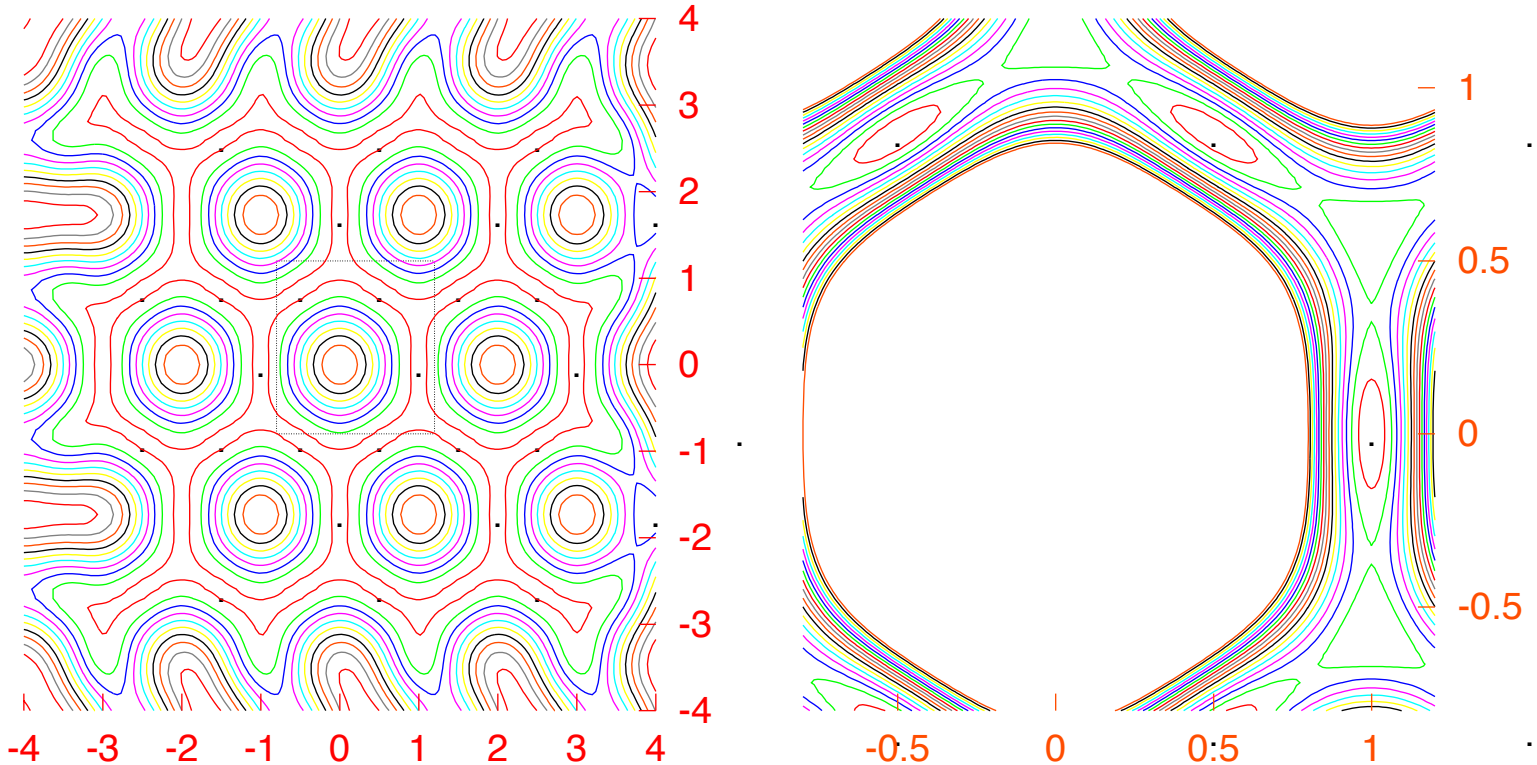
probability of observed mass of a randomly uniformly picked butterfly

$$p(x) = \frac{2}{5}\phi(x; -1.25, 1) + \frac{3}{5}\phi(x; 2.95, 1)$$

Mixing Gaussian distribution together

Bumps and More Bumps (MacKay and Williams)

A mixture of k Gaussians models can have $\frac{5}{3}k$ modes.



Mixtures

- *Mixture of f and g*:

$$p(x) = \eta f(x) + (1 - \eta) g(x)$$

  Simplest, most common kind of latent variable model

- *Hidden variable represention*: Define $Z \sim \text{Bernoulli}(\eta)$ and

$$p(x) = \sum_{z=0,1} p(x \mid z)\, p(z)$$

  with $p(x \mid 1) = f(x)$, $p(x \mid 0) = g(x)$, $p(z) = \eta^z (1 - \eta)^{(1-z)}$.

Bayesian Inference

underlying para θ

The parameter θ of a model is viewed as a random variable.
Inference usually carried out as follows:

- Choose a *generative model* $p(x \mid \theta)$ for the data.
- Choose a *prior distribution* $\pi(\theta)$ that expresses beliefs about the parameter before seeing any data.
- After observing data $\mathcal{D}_n = \{x_1, \ldots, x_n\}$, update beliefs and calculate the *posterior distribution* $p(\theta \mid \mathcal{D}_n)$.

Bayes' Theorem

use this to connect $p(x|\theta)$ $\pi(\theta)$ $p(\theta|D(x_1, \dots x_n))$

$$\begin{aligned}\mathbb{P}(A|B) &= \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \\ &= \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)}\end{aligned}$$

Bayes' Theorem

The posterior distribution can be written as

$$p(\theta \mid x_1, \ldots, x_n) = \frac{p(x_1, \ldots, x_n \mid \theta)\pi(\theta)}{p(x_1, \ldots, x_n)} = \frac{\mathcal{L}_n(\theta)\pi(\theta)}{c_n} \propto \mathcal{L}_n(\theta)\pi(\theta)$$

where $\mathcal{L}_n(\theta) = \prod_{i=1}^{n} p(x_i \mid \theta)$ is the *likelihood function* and

$$c_n = p(x_1, \ldots, x_n) = \int p(x_1, \ldots, x_n \mid \theta)\pi(\theta)d\theta = \int \mathcal{L}_n(\theta)\pi(\theta)d\theta$$

is the normalizing constant, which is also called *evidence*.

$$\int \frac{\mathcal{L}_n(\theta)\,\pi(\theta)}{c_n} d\theta = 1$$

Example

$X \sim$ Bernoulli(θ) with data $\mathcal{D}_n = \{x_1, \ldots, x_n\}$.

$$\mathbb{P}(X_i = 1|\theta) = \theta^{x_i}(1-\theta)^{1-x_i}$$

Natural prior distribution on θ: Beta(α, β) distribution

$$\pi_{\alpha,\beta}(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

Example

$X \sim$ Bernoulli(θ) with data $\mathcal{D}_n = \{x_1, \ldots, x_n\}$.

$$\mathbb{P}(X_i = 1|\theta) = \theta^{x_i}(1-\theta)^{1-x_i}$$

Natural prior distribution on θ: Beta(α, β) distribution

conjugate prior 共轭先验

$$\pi_{\alpha,\beta}(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

$$\pi_{\alpha,\beta}(\theta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{\int_{\theta=1}^{1} \theta^{\alpha-1}(1-\theta)^{\beta-1}\mathrm{d}\theta}$$

$$\pi_{\alpha,\beta}(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

normalizing constant

$$\pi_{\alpha,\beta}(\theta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{\int_{\theta=1}^{1} \theta^{\alpha-1}(1-\theta)^{\beta-1}\mathrm{d}\theta}$$

$$\pi_{\alpha,\beta}(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

gamma function $$\Gamma(z) = \int_{x=0}^{\infty} x^{z-1}e^{-x}\mathrm{d}x$$

Example

$X \sim$ Bernoulli(θ) with data $\mathcal{D}_n = \{x_1, \ldots, x_n\}$. Natural prior distribution on θ: Beta(α, β) distribution

$$\pi_{\alpha,\beta}(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

Let $s = \sum_{i=1}^{n} x_i$ be the number of "successes."

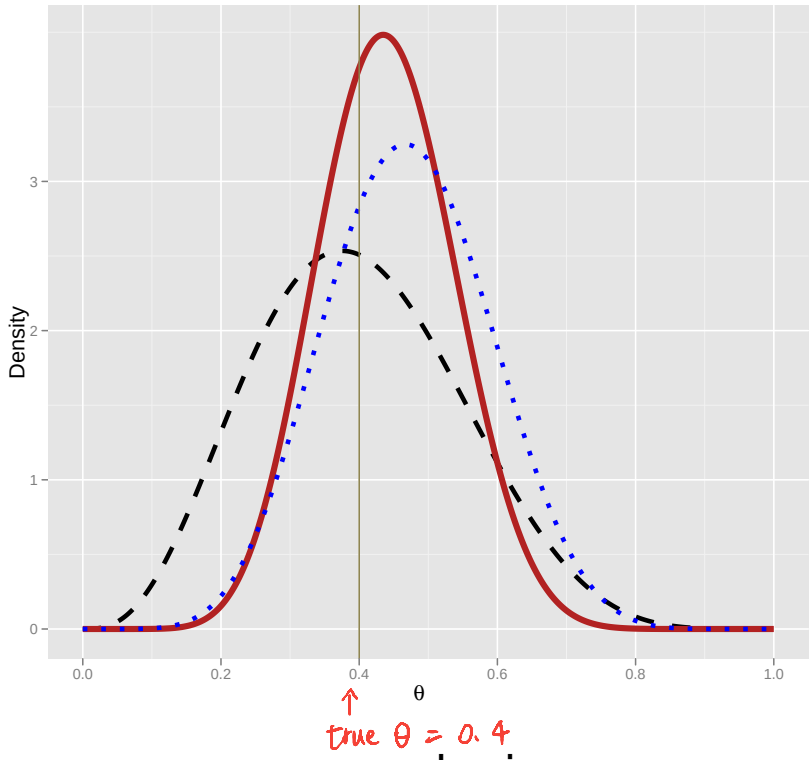
Posterior distribution $\theta \mid \mathcal{D}_n$ is Beta($\alpha + s, \beta + n - s$). Posterior mean is

$$\bar{\theta} = \frac{\alpha + s}{\alpha + \beta + n} = \left(\frac{n}{\alpha + \beta + n}\right)\widehat{\theta} + \left(\frac{\alpha + \beta}{\alpha + \beta + n}\right)\theta_0$$

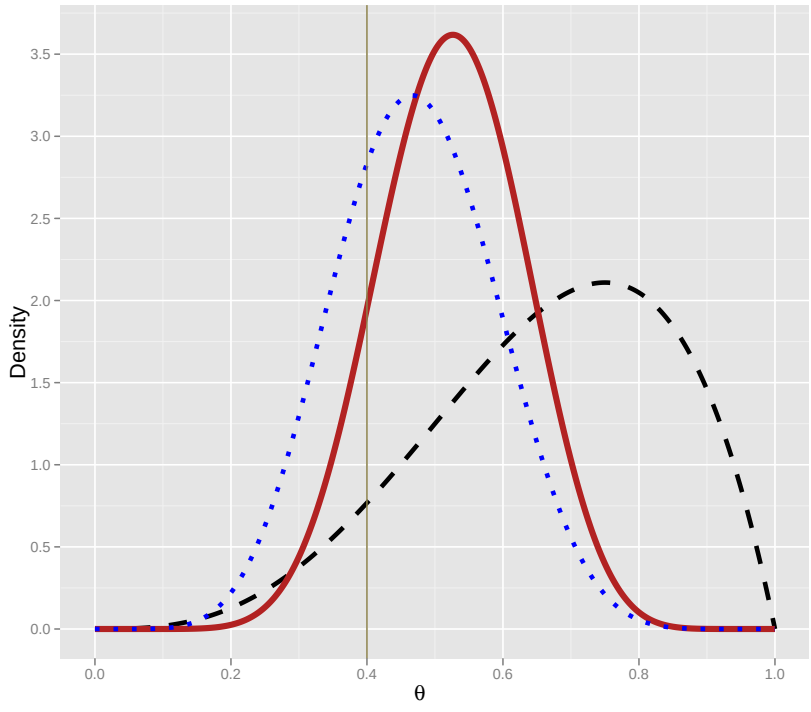
where $\widehat{\theta} = s/n$ is the MLE and $\theta_0 = \alpha/(\alpha + \beta)$ is the prior mean. This is an example of Bayesian shrinkage (connections to regularization)

Example

$n = 15$ points sampled as $X \sim \text{Bernoulli}(\theta = 0.4)$, with $s = 7$ heads.



good prior



bad prior

Prior distribution (black-dashed), likelihood function (blue-dotted), posterior distribution (red-solid).

↑ we try to maximize

Dirichlet and Multinomial

Multinomial model with Dirichlet prior is generalization of the Bernoulli/Beta model.

param is α, random variable is θ

$$\text{Dirichlet}_\alpha(\theta) = \underbrace{\frac{\Gamma(\sum_{j=1}^K \alpha_j)}{\prod_{j=1}^K \Gamma(\alpha_j)}}_{\text{Normalizing constant}} \prod_{j=1}^K \theta_j^{\alpha_j - 1}$$

where $\alpha = (\alpha_1, \ldots, \alpha_K) \in \mathbb{R}_+^K$ is a non-negative vector.

Data $x_i \in [K]$

multinomial model: param is θ, random variable is x

$$\mathbb{P}(x_i|\theta) = \prod_{k=1}^K \theta_k^{\mathbb{1}(x_i=k)}$$

$p(x_i = k|\theta) = \theta_k$

Bayesian Mixtures

Mixture of k Gaussians $p(x) = \sum_{j=1}^{k} \eta_j \phi(x; \mu_j, \sigma_j^2)$.

mean variance

Let $z_i = (z_{i1}, \ldots, z_{ik})$ be random vector of length k where $z_{ij} = 1$ if x_i came from the jth component $\phi(x; \mu_j, \sigma_j^2)$. Common choice of priors:

every param can be random ⇒ can put a prior for every param

$$
\begin{aligned}
\xi &\sim \text{Dirichlet}(\beta) \text{ where } \beta \in \mathbb{R}_+^k \\
z_1, \ldots, z_n \mid \xi &\sim \text{Multinomial}(\xi) \\
\sigma_1^2, \ldots, \sigma_k^2 &\sim \text{Inverse-Gamma}(\nu_0, \sigma_0^2) \\
\mu_j \mid \sigma_j &\sim \text{Normal}(\mu_0, \sigma_j^2)
\end{aligned}
$$

Standard Bayesian techiques then infer posterior distribution of $\mu_1, \ldots, \mu_k$, $\sigma_1^2, \ldots, \sigma_k^2$, and $\eta_1, \ldots, \eta_k$ given the data $x_1, \ldots, x_n$, by simulating the membership indicators z_{ij}.

採樣

Summary

- Mixtures are latent variable models
- The mixing weight encodes a hidden variable
- Computing with mixtures uses basic probabilistic reasoning
- But can get complicated
- Topic models are flexible mixtures models for complex data like documents and images (next)