LeBron James : $\{0, 1, 0, 0, 1, 1, 0, 1, 1, 0, 0, 1, 1, 1, 1, 0, 0, 1, \ldots\}$

exchangeable data model (EDM)

binary data

$x_i \in \{0, 1\}$

dataset: $x_{1:n}$

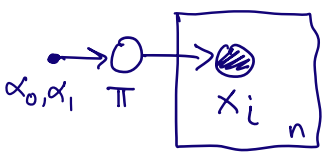
assume:

1) each shot is indpt drawn from $p(x|\pi)$

2) the parameter comes from a prior.

$$p(\pi, x) = p(\pi) \prod_{i=1}^{n} p(x_i|\pi)$$

prior — likelihood



$p(\pi | x_{1:n})$

Bernoulli:

$$p(x|\pi) = \pi^x (1-\pi)^{(1-x)} \qquad \pi \in [0, 1]$$

$$\mathbb{E}[X|\pi] = \pi$$

$$\text{Var}(X|\pi) = \pi(1-\pi)$$

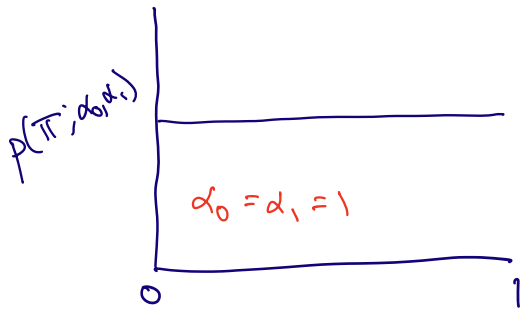
Beta

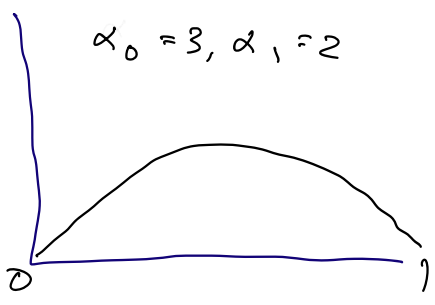
normalizing constant

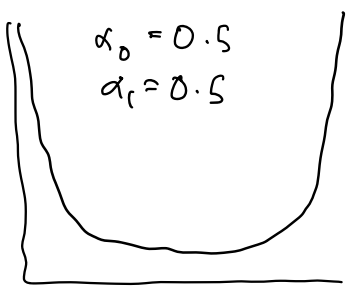
$$p(\pi; \alpha_0, \alpha_1) = \frac{\Gamma(\alpha_1 + \alpha_0)}{\Gamma(\alpha_1)\Gamma(\alpha_0)} \pi^{\alpha_1 - 1} (1-\pi)^{\alpha_0 - 1}$$

$\alpha_0, \alpha_1 > 0$

$\pi \in (0, 1)$







$$\mathbb{E}[\pi; \alpha_1, \alpha_0] = \frac{\alpha_1}{\alpha_1 + \alpha_0}$$

$$\mathbb{E}[1-\pi; \alpha_1, \alpha_0] = \frac{\alpha_0}{\alpha_1 + \alpha_0}$$

$$\mathrm{Var}(\pi; \alpha_1, \alpha_0) = \alpha_1 \alpha_0 \Big/ (\alpha_1 + \alpha_0)^2 (\alpha_1 + \alpha_0 + 1)$$

$x_{1:n}$: binary dataset

$$p(\pi) \prod_{i=1}^{n} p(x_i \mid \pi)$$

Beta ↑ Bernoulli ↑

$$\pi \sim \mathrm{Beta}(\alpha_0, \alpha_1)$$

$$x_i \sim \mathrm{Bern}(\pi) \quad i = 1..n$$

$$p(a \mid b) \propto p(a, b)$$

$$p(\pi \mid x_{1:n}; \alpha_0, \alpha_1) \propto p(\pi; \alpha_0, \alpha_1) \prod_{i=1}^{n} p(x_i \mid \pi)$$

$$= \pi^{\alpha_1 - 1} (1-\pi)^{\alpha_0 - 1} \prod_{i=1}^{n} \pi^{x_i} (1-\pi)^{(1-x_i)}$$

$$= \pi^{\left(\alpha_1 + \sum_{i=1}^{n} x_i - 1\right)} (1-\pi)^{\left(\alpha_0 + \sum (1-x_i) - 1\right)}$$

This is a beta distribution.

$$\hat{\alpha}_1 = \alpha_1 + \sum_{i=1}^{n} x_i \quad \leftarrow n_1$$

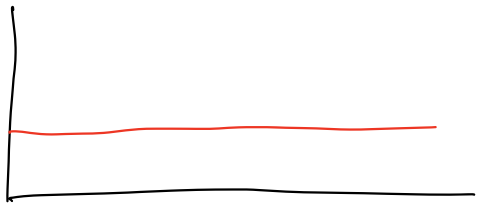
$$\hat{\alpha}_0 = \alpha_0 + \sum_{i=1}^{n} (1 - x_i) \quad \leftarrow n_0$$

} posterior beta

Conjugacy – posterior is in the same family as the prior.

Beta / Bernoulli are called conjugate pair

$$\mathbb{E}[\pi \mid x_{1:n}] = \frac{\alpha_1 + n_1}{\alpha_1 + \alpha_0 + n_1 + n_2} = \frac{\alpha_1 + n_1}{\alpha_1 + \alpha_0 + n}$$



joint : $p(\pi, x_{1:n})$

posterior : $p(\pi \mid x_{1:n})$

predictive : $p(x' \mid x_{1:n}) = \int p(\pi, x' \mid x_{1:n})\, d\pi$

$$= \int p(x' \mid \pi, x_{1:n})\, p(\pi \mid x_{1:n})\, d\pi$$

$$= \int p(x' \mid \pi)\, p(\pi \mid x_{1:n})\, d\pi$$

$$= \int \pi\, p(\pi \mid x_{1:n})\, d\pi$$

$$= \mathbb{E}[\pi(x_{1:n}]$$

Categorical data
Count data
Real-valued data

---

x_i takes on one of K "categories" e.g., terms in a vocabulary.

Example: Text.

"To be or not to be, that is the question."

x_1 x_2 x_3 x_4 x_5

indicator vector, "one-hot" vector $x \in \mathbb{1}^K$

a	0
of	0
thou	0
Juliet	0
Romeo	0
be	1
⋮	0

K

Categorical distribution

parameter θ — prob of each value

$$\theta = \begin{pmatrix} 0.01 \\ 0.02 \\ 0.0 \\ 0.005 \\ \vdots \end{pmatrix} \quad K \qquad \sum_{k=1}^{K} \theta_k = 1 \quad \theta_k \geq 0$$

$K-1$ simplex, $\theta \in \Delta^{K-1}$

$$p(x \mid \theta) = \prod_{k=1}^{K} \theta_k^{x_k}$$

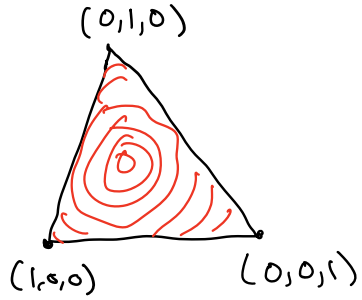
$p(x \mid \theta) = \prod_{k=1}^{K} \theta_k^{x^{(k)}}$ $\quad x^{(k)} \left(\text{Bern} \quad \pi^x (1-\pi)^{1-x} \right)$

$\mathbb{E}[x^{(k)} \mid \theta] = \theta_k$

$\text{Var}[x^{(k)} \mid \theta] = \theta_k (1 - \theta_k)$

$\text{Cov}[x^{(j)}, x^{(k)} \mid \theta] = -\theta_j \theta_k$

$$p(\theta, x_{1:n}) = p(\theta) \prod_{i=1}^{n} p(x_i \mid \theta)$$



$$p(\theta; \alpha) = \left(\frac{\Gamma(\Sigma_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \right) \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}$$

$\alpha = (\alpha_1, \ldots, \alpha_K), \alpha_k > 0$

$$\mathbb{E}[\theta_j; \alpha] = \frac{\alpha_j}{\sum_{k=1}^{K} \alpha_k} \qquad \text{Var}[\theta_j; \alpha] = \frac{\alpha_j \left(\sum_{k \neq j} \alpha_k \right)}{\left(\sum_k \alpha_k \right)^2 \left(\sum_k \alpha_k + 1 \right)}$$

$$\text{Cov}[\theta_j, \theta_\ell; \alpha] = \frac{-\alpha_j \alpha_\ell}{\quad}$$

Exchangable Dirichlet

$p(\theta; \alpha_0) : \text{Dir}(\alpha_0, \alpha_0, \ldots, \alpha_0)$ (k times)

$\alpha_0 > 0$

$x: \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix} \quad \alpha: \begin{pmatrix} 3 \\ 2 \\ 9.6 \\ 0.5 \end{pmatrix}$

$\theta: \begin{pmatrix} p(x=1) \\ p(x=2) \\ p(x=3) \\ p(x=4) \end{pmatrix}$

$\theta \sim \text{Dir}_k(\alpha_0)$

$x_i \sim \text{Cat}(\theta)$

$p(\theta \mid x_{1:n}) : \text{Dirichlet}(\hat{\alpha}_1, \ldots, \hat{\alpha}_K)$

$$\hat{\alpha}_k = \alpha_k + \sum_{i=1}^{n} x_i^{(k)}$$

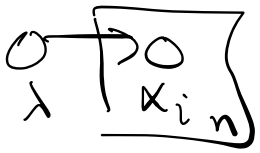
$\sum_i x_i^{(k)}$

$$\mathbb{E}[\theta_k \mid x_{1:n}, \alpha] = \frac{\alpha_k + n_k}{\sum \alpha_j + n}$$ Laplace smoothing.

---

daily clicks: $\{0, 5, 3, 1, 9, 6, \ldots\}$ count data.

$$p(\lambda, x_{1:n}) = p(\lambda) \prod_{i=1}^{n} p(x_i \mid \lambda)$$



Poisson distribution

$$p(x \mid \lambda) = \frac{\lambda^x \exp\{-\lambda\}}{x!}$$

$\lambda \geq 0$

$$\mathbb{E}[x \mid \lambda] = \lambda \qquad \mathrm{Var}[x \mid \lambda] = \lambda$$

$$p(\lambda; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \exp\{-\beta\lambda\} \lambda^{\alpha-1}$$ Gamma prior

$\alpha > 0, \beta > 0$

$\alpha = 1$: exponential dist.

$$\mathbb{E}[\lambda] = \alpha/\beta \qquad \mathrm{Var}[\lambda] = \alpha/\beta^2$$

$p(\lambda \mid x_{1:n}, \alpha, \beta)$: Gamma$(\hat{\alpha}, \hat{\beta})$

$$\hat{\alpha} = \alpha + \sum_{i=1}^{n} x_i$$
$$\hat{\beta} = \beta + n$$

$$\mathbb{E}[\lambda \mid x_{1:n}, \alpha, \beta] = \frac{\alpha + \sum_{i=1}^{n} x_i}{\beta + n}$$

$\mathbb{E}[\lambda; \hat{\alpha}, \hat{\beta}]$

Gaussian.

$p(x \mid \mu, \sigma^2)$ $\quad x \in \mathbb{R}$ $\quad \sigma^2 > 0$, $\mu \in \mathbb{R}$

$$p(\mu, x_{1:n}) = p(\mu) \prod_{i=1}^{n} p(x_i \mid \mu)$$

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

$$\mu \sim N(\mu_0, \lambda^2)$$

$$x_i \mid \mu \sim N(\mu, \sigma^2)$$

$$\mu \mid x_{1:n} \sim N(\hat{\mu}_0, \hat{\lambda}^2) \qquad \bar{x} = \sum_i x_i / n$$

$$\hat{\mu}_0 = \left(\frac{\mu_0/\lambda^2 + n\bar{x}/\sigma^2}{1/\lambda^2 + n/\sigma^2} \right)$$

---

hyperparameters. parameters to the prior.

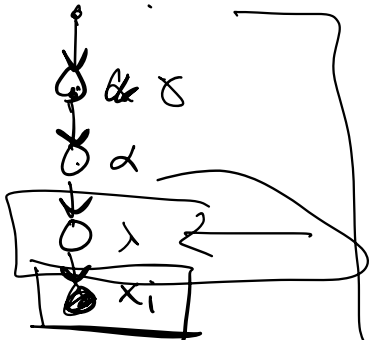
"Bayesian purist"

"Orthodox Bayesian"

"Crayesians"

Uniform prior.

"Vague prior"



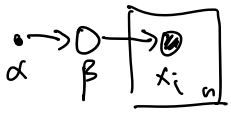
Model Criticism.

held out by likelihood.

$\beta \sim p(\beta; \alpha)$

$x_i \sim p(x \mid \beta)$

generative process



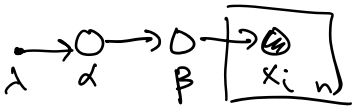
graphical model

$$p(\beta, x_{1:n}) = p(\beta; \alpha) \prod_{i=1}^{n} p(x_i \mid \beta)$$

factorized joint

$p(\beta \mid x_{1:n}; \alpha)$

posterior



$p(x_{new} \mid x_{1:n}; \alpha)$

posterior predictive

Held out log probability

$$p(\theta, x_{1:n}) = p(\theta; \alpha_0) \prod_{i=1}^{n} p(x_i \mid \theta)$$

$p(\theta; \alpha_0)$: categorical dist parameter over terms

$p(x_i \mid \theta)$: categorical dist w/ param θ

$$p(x_{new}^{=v} \mid x_{1:n}; \alpha_0) = \int \underbrace{p(x_{new}^{=v} \mid \theta)}_{\theta_v} p(\theta \mid x_{1:n}; \alpha_0)\, d\theta$$

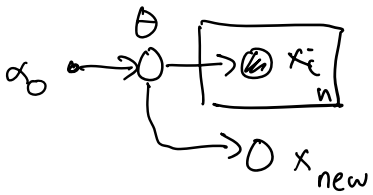
x_{new}: a word

θ: dist. over words, Δ^{V-1}

$$= \mathbb{E}[\theta_v \mid x_{1:n}; \alpha_0]$$

$\hat{\alpha}_v = \alpha_0 + n_v$ ← # times Shakespeare used term v

$$\mathbb{E}[\theta_v \mid x_{1:n}; \alpha_0] = \frac{\alpha_0 + n_v}{V\alpha_0 + n}$$



$\mathbb{E}[\theta \mid x_{in}] \triangleq \hat{\theta}$

data splitting

$x_{in,\ 1:n_{in}}$

$x_{out,\ 1:n_{out}}$

$$\hat{\Lambda}(\alpha_0) \triangleq \frac{1}{n_{out}} \sum_{i=1}^{n_{out}} \log p(x_{out,i} \mid x_{in}; \alpha_0)$$

Model selection.

prior predictive checks (Box)
posterior " " (Rubin ...)

$$\hat{\alpha}_0 = \arg\max_{\alpha_0} \log p(x_{1:n}; \alpha_0)$$

$$\int p(x_{1:n} \mid \theta)\, p(\theta)\, d\theta$$

$F(x)$ — true population distribution of a datapoint.

$$\Lambda(\alpha_0) \triangleq \mathbb{E}_F\left[\log p(X_{new} \mid x_{1:n}; \alpha_0)\right]$$

"proper scoring function"

$$\hat{\Lambda}(\alpha_0) \triangleq \frac{1}{n_{out}} \sum_{i=1}^{n_{out}} \log p(x_{i,out} \mid x_{in}; \alpha_0)$$

Conditional models

Linear regression
Logistic regression
Stochastic optimization for MAP estimation.

$$y_i \mid x_i, \beta \sim \mathcal{N}(\beta \cdot x_i, \sigma^2)$$

x_i: p-vector of features; β: p-vector of coefficients; $=$

$$p(\beta, y_{1:n} \mid x_{1:n}; \lambda) = p(\beta; \lambda) \prod_{i=1}^{n} p(y_i \mid x_i, \beta)$$

$$\beta_k \sim \mathcal{N}(0, \lambda^2)$$

$$p(\beta \mid y_{1:n}, x_{1:n}, \lambda)$$

$$\hat{\beta}_{MAP} = \arg\max_{\beta} p(\beta \mid y_{1:n}, x_{1:n}, \lambda)$$

$$\hat{\beta}_{MAP} = \arg\max_{\beta} \left(\log p(\beta; \lambda) + \sum_{i=1}^{n} \log p(y_i \mid x_i, \beta) \right)$$

Generative process:

For each coefficient k:

$$\beta_k \sim N(0, \lambda^2) \quad \text{prior}$$

For each data point i:

$$y_i \mid x_i, \beta \sim N(\beta \cdot x_i, \sigma^2)$$

$$y_i = \beta \cdot x_i + \varepsilon_i$$
$$\varepsilon_i \sim N(0, \sigma^2)$$

$$p(\beta \mid \mathcal{D}; \lambda) = \frac{p(\beta) \prod p(y_i \mid x_i, \beta)}{p(y_{1:n} \mid x_{1:n})}$$

(x_i, y_i)
$i = 1 \ldots n$

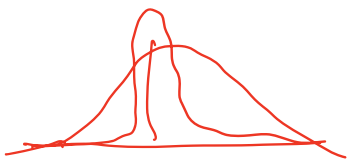
$$p(\beta \mid y_{1:n}, x_{1:n}, \lambda)$$

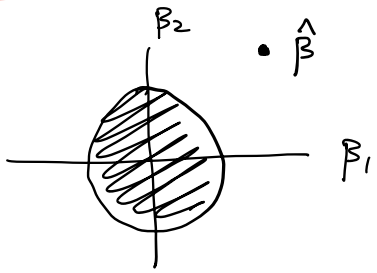
$$\hat{\beta}_{MAP} = \arg\max_{\beta} \log p(\beta) + \sum_{i=1}^{n} \log p(y_i \mid x_i, \beta) - \log p(y_{1:n} \mid x_{1:n})$$

$$= \arg\max_{\beta} \underbrace{\log p(\beta) + \sum_{i=1}^{n} \log p(y_i \mid x_i, \beta)}_{\text{log joint}} + \text{const.}$$

Ridge regression.

$$= \arg\max_{\beta} -\frac{1}{2\lambda^2} \sum_{k=1}^{p} \beta_k^2 - \sum_{i=1}^{n} \frac{(y_i - \beta \cdot x_i)^2}{2\sigma^2} + \text{const.}$$





Laplace prior or lasso.

$$\beta_k \sim \text{Laplace}(\lambda) \qquad k = 1..p$$

$$y_i \mid x_i, \beta \sim N(\beta \cdot x_i, \sigma^2) \qquad i = 1..n$$

$$p(\beta_k; \lambda) = \frac{1}{2\lambda} \exp\left\{ \frac{-|\beta_k|}{\lambda} \right\}$$

$$\ell_{MAP}(\beta) \triangleq -\frac{1}{\lambda} \sum_{k=1}^{p} |\beta_k| + \sum_{i=1}^{n} \log p(y_i \mid x_i, \beta)$$

MAP estimate tends to be sparse

---

movie review data (x_i, y_i) $i = 1..n$

$x_i : (x_i^{(1)}, \ldots, x_i^{(p)})$ word counts across a vocabulary.

$y_i : \{0, 1\}$ $\quad 0 = \text{bad}$, $1 = \text{good}$

$$\beta_k \sim N(0, \lambda^2) \qquad k = 1..p$$

$$y_i \mid x_i, \beta \sim \text{Bern}(\sigma(\beta \cdot x_i))$$

$\sigma(\omega)$: sigmoid function $\quad \mathbb{R} \to (0, 1)$

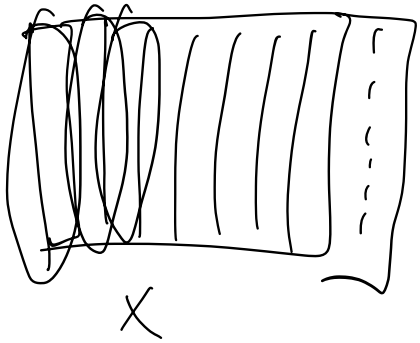
$$\sigma(\beta \cdot x_i) = \frac{1}{1 + \exp\{-\beta \cdot x_i\}}$$

$$-\frac{1}{2\sigma^2} q^2$$

$$\hat{\beta}_{MAP} = \arg\max_{\beta} \; -\frac{1}{2\lambda^2} \sum_{k=1}^{p} \beta_k^2 + \sum_{i=1} y_i \log \sigma(\beta \cdot x_i) + (1 - y_i) \log \sigma(-\beta \cdot x_i)$$

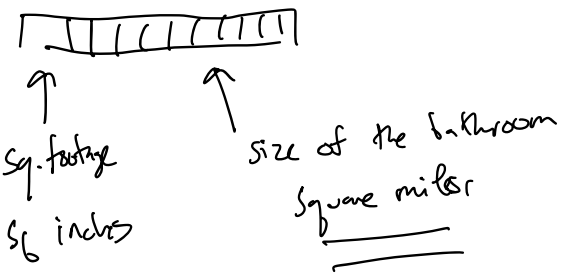
$$p(z) = \pi^z (1-\pi)^{1-z} \qquad 1 - \sigma(\omega) = \sigma(-\omega)$$

$p(\beta \mid y_{1:n}, x_{1:n}, \lambda)$



① Standardizing the covariates

$x_i, y_i \quad i = 1 \dots n$



② The intercept. $y_i \sim N(\underbrace{\beta \cdot x_i}_{=} + \underbrace{\alpha}_{=}, \sigma^2)$

$\beta_k \sim N(0, \lambda^2)$

$\alpha \sim N(0, \gamma^2)$ ← γ^2 should be big.

Setting the regularizer.

$\mathbb{E}[Y \mid x^*, \hat{\beta}_{MAP}] = \hat{\beta}_{MAP} \cdot x^*$ linear regression

$\mathbb{E}[Y \mid x^*, \hat{\beta}_{MAP}] = \sigma(\hat{\beta}_{MAP} \cdot x^*)$ logistic regression

$(x_{in}, y_{in}) \quad (x_{out}, y_{out})$

$$\hat{\Lambda}(\lambda) = \frac{1}{m} \sum_{i=1}^{m} \log p(y_{i,out} \mid x_{i,out}, \hat{\beta}(\lambda))$$

$\hat{\beta}(\lambda)$ = MAP estimate with hyperparameter λ.

$\beta_k \sim N(0, \lambda^2) \quad k = 1 \cdots p$

$y_i | x_i, \beta \sim p(y | x_i, \beta) \quad i = 1..n$

$\hat{\beta}_{MAP} = \arg\max_{\beta} \log p(\beta; \lambda) + \sum_{i=1}^{n} \log p(y_i | x_i, \beta)$

$\Lambda(\lambda) \triangleq \mathbb{E}_F[\underbrace{\log p(Y | X, \mathcal{D})}_{\text{posterior predictive}}]$ $\quad x', y' \sim F$, $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$ — observed

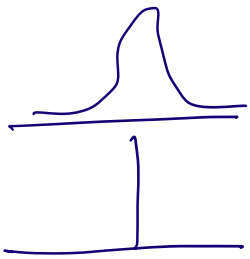
$\hat{\Lambda}(\lambda) \approx \frac{1}{n_{out}} \sum_{i=1}^{n_{out}} \log p(y_{i,out} | x_{i,out}, \mathcal{D}_{in}) \quad \mathcal{D}_{in} = \{(x_{in,i}, y_{in,i})\}_{i=1}^{n_{in}}$

$$p(y | x, \mathcal{D}) = \int p(y | x, \beta)\, p(\beta | \mathcal{D})\, d\beta$$

$$\Downarrow$$

$$\delta(\beta_{MAP})$$

$$\approx p(y | x, \hat{\beta}_{MAP}(\mathcal{D}))$$



$$\hat{\Lambda}(\lambda) \approx \frac{1}{n_{out}} \sum_{i=1}^{n_{out}} \log p(y_{i,out} | x_{i,out}, \hat{\beta}^{\lambda}_{MAP}(\mathcal{D}_{in}))$$

$$\equiv \hat{\Lambda}(\lambda) = \sum_{i=1}^{n_{out}} -(y_{i,out} - \hat{\beta}^{\lambda}_{MAP} \cdot x_{i,out})^2 / 2\sigma^2$$

$$\equiv \hat{\Lambda}(\lambda) = \sum_{i=1}^{n_{out}} y_{i,out} \log \sigma(\hat{\beta}_{MAP} \cdot x_{out}) + (1 - y_{i,out}) \log \sigma(-\hat{\beta}_{MAP} \cdot x_{out})$$

How do we find β_{MAP}?

In: $\mathcal{D}$ Out: β_{MAP}

$$\mathcal{L}(\beta) = \log p(\beta; \lambda) + \sum_{i=1}^{n} \log p(y_i | x_i, \beta)$$

gradient ascent.

$$\nabla_\beta \ell(\beta) = \left(\frac{\partial}{\partial \beta_1} \ell(\beta), \ldots, \frac{\partial}{\partial \beta_p} \ell(\beta) \right)$$

initialize $\beta[0]$

$$\beta[t+1] = \beta[t] + \rho_t \nabla_\beta \ell(\beta[t])$$

↑ "step size", "learning rate", 0.001

LINEAR REGRESSION

$$\ell(\beta) = -\frac{1}{2\lambda^2} \sum_{k=1}^{p} \beta_k^2 - \sum_{i=1}^{n} \frac{(y_i - \beta \cdot x_i)^2}{2\sigma^2}$$

$$\frac{\partial}{\partial \beta_k} \ell(\beta) = -\beta_k / \lambda^2 + \sum_{i=1}^{n} \frac{(y_i - \beta \cdot x_i)}{\sigma^2} x_{i,k}$$

$$\boxed{\nabla_\beta \ell(\beta) = -\frac{1}{\lambda^2} \beta + \sum_{i=1}^{n} \frac{(y_i - \beta \cdot x_i)}{\sigma^2} x_i}$$

signed residual $\triangleq (y_i - \mathbb{E}[Y | x_i, \beta])$

$= (y_i - \beta \cdot x_i)$

Logistic regression

$$\ell(\beta) = \frac{-1}{2\lambda^2} \sum \beta_k^2 + \sum_{i=1}^{n} \left(y_i \log \sigma(\beta \cdot x_i) + (1 - y_i) \log \sigma(-\beta \cdot x_i) \right)$$

$$\frac{d}{d\eta} \sigma(\eta) = \sigma(\eta)(1 - \sigma(\eta))$$

$$\nabla \ell(\beta) = -\beta / \lambda^2 + \sum_{i=1}^{n}$$

$$\frac{d}{d\eta_i} \log p(y_i | x_i, \beta) = \frac{y_i \, \sigma(\eta_i)(1 - \sigma(\eta_i))}{\sigma(\eta_i)} + \frac{(1 - y_i) \, \sigma(-\eta_i)(1 - \sigma(-\eta_i))}{\sigma(-\eta_i)}$$

$\eta_i \triangleq \beta \cdot x_i$

$$= y_i (1 - \sigma(\eta_i)) - (1 - y_i)(1 - \sigma(-\eta_i))$$

$$= y_i (1 - \sigma(\eta_i)) - (1 - y_i)(\sigma(\eta_i))$$

$$= y_i - y_i \sigma(\eta_i) - \sigma(\eta_i) + y_i \sigma(\eta_i)$$

$$= y_i - \sigma(\eta_i)$$

$$\nabla_\beta \ell(\beta) = -\frac{1}{\lambda^2}\beta + \sum_{i=1}^{n} (y_i - \sigma(\beta \cdot x_i))\, x_i$$

$$\varepsilon_i = y_i - \mathbb{E}[Y \mid x_i, \beta] = \sigma(\beta \cdot x_i)$$

$$\nabla_\beta \ell(\beta) = -\frac{1}{\lambda^2}\beta + \sum_{i=1}^{n} (y_i - \mathbb{E}[Y \mid x_i, \beta])\, x_i$$

Stochastic optimization

10M more reviews + sentiment.

large data set $y_{1:n}$

$$\ell(\beta; y_{1:n}) = h(\beta) + \sum_{i=1}^{n} f(\beta; y_i)$$

$$\hat{\beta} = \arg\max_{\beta} \ell(\beta; y_{1:n})$$

$$\nabla_\beta \ell = \nabla_\beta h(\beta) + \sum_{i=1}^{n} \nabla_\beta f(\beta; y_i)$$

stochastic gradient ascent.

Robbins + Monro (1951), Bottou (1998)

Stochastic gradient:

$$g \sim p(g; \beta)$$

$$\mathbb{E}[G] = \nabla_\beta \ell(\beta; y_{1:n})$$ unbiased stochastic gradient

repeat ~~forever~~ until convergence

$$g_t \sim p(g; \beta[t])$$

$$\beta[t+1] = \beta[t] + \rho_t g_t$$

ρ_t step size schedule

Robbins Monro conditions $\frac{1}{t}$

$$\sum_t \rho_t = \infty \qquad \sum_t \rho_t^2 < \infty$$

$\hat{F}$: empirical dist of data $1/n$ on each data point

$$y \sim \hat{F}$$

$$g = \nabla_\beta h(\beta) + n \nabla_\beta f(\beta[t]; y)$$

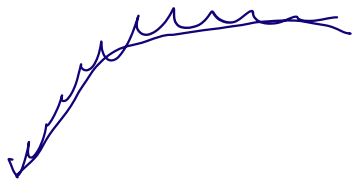
$$\mathbb{E}_{\hat{F}}[G] = \nabla \ell(\beta) \quad \text{— confirm.}$$

Minibatch $\quad g = \nabla_\beta h(\beta) + \frac{n}{B} \sum_{b=1}^{B} \nabla_\beta f(\beta[t]; y_b)$

B batch size

$y_1 \cdots y_B \sim \hat{F}$

Practice



1. Convergence.
   - norm of the noisy gradient
   - objective plateaued
   - validation objective
2. Minibatches + data sweeps
   w/o replacement
3. Step size Schedule
   - Robbins Monro
   - Constant step size (η)
   - Adaptive step sizes — Ada grad (η)
     ADAM (η)
     RMS Prop (η)
     Ada delta (η)

$p > n$

- Mixture models, mixed membership models, factor models
- Gibbs sampling (MCMC), variational inference.

Mixture model

600 flowers

group the data
1. we don't know which data belongs to which group
2. we don't know what the clusters are.

$x_i : (x_{i,1}, \ldots, x_{i,d})$ — d dimensions, $d = 512$

Core assumption: each x_i belongs to a cluster, drawn from a distribution associated with it.

generative process, etc.

observed
- n data points $x_1 \cdots x_n$

K — # mixture components.

latent
- n assignments $z_1 \cdots z_n$
  z_i is a K-categorical variable, $z_i \in \mathbb{1}^K$
- K mixture components, $\beta_1 \cdots \beta_K$
  parameterizes a dist over x
- proportions $\theta \in \Delta^K$

1. generative process.
2. factorized joint

$$p(\theta, \beta_{1:K}, z_{1:n}, x_{1:n}) = p(\theta) \prod_{k=1}^{K} p(\beta_k) \prod_{i=1}^{n} \left(p(z_i \mid \theta) \, p(x_i \mid z_i, \beta_{1:K}) \right)$$

$p(\theta)$: Dirichlet —

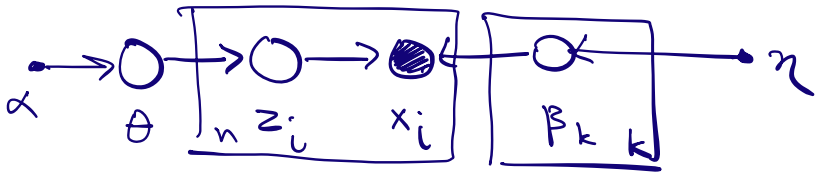
$p(\beta_k)$: K Gaussians —

$p(z_i \mid \theta)$: Categorical (θ)

$$p(x_i \mid z_i, \beta_{1:K}) = \prod_{k=1}^{K} f(x_i ; \beta_k)^{z_i^{(k)}} \qquad f_{\beta_k}(x_i)$$

$$f(x_i ; \beta) = \prod_{j=1}^{d} \ell(x_{id} ; \beta^{(d)})$$

(3)



$$p(\theta, \beta_{1:K}, z_{1:n} \mid x_{1:n})$$

GOAL, Assumptions. DATA

↓ ↓

MODEL → POSTERIOR → USE IT

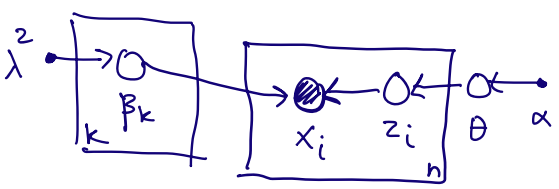
10/5 Goal: Cluster data

Mixture model:

$\theta \sim \text{Dirichlet}(\alpha)$

For each component k: (global variables)

$\beta_k \sim N_d(0, \lambda^2)$



For each data point i: (local variables)

$z_i \mid \theta \sim \text{Cat}(\theta)$

$x_i \mid z_i, \beta_{1:K} \sim N(\beta_{z_i}, \sigma^2)$

$$p(\theta, \beta_{1:K}, z_{1:n}, x_{1:n}; \lambda, \alpha) = p(\theta) \prod_k p(\beta_k) \prod_i p(z_i \mid \theta)\, p(x_i \mid z_i, \beta_{1:K})$$

hidden: $\theta, \beta_{1:K}, z_{1:n}$ observed: $x_{1:n}$

local latent variable: z_i

posterior mixture

$$p(\theta, \beta_{1:K}, z_{1:n} \mid x_{1:n})$$

$$\log p(\theta, \beta_{1:K}, z_{1:n} \mid x_{1:n}) = \log p(\theta, \beta_{1:K}, z_{1:n}, x_{1:n}) - \log p(x_{1:n})$$

$$\log p(-) = \log p(\theta) + \sum_{k=1}^{K} \log p(\beta_k) + \sum_{i=1}^{n} \log p(z_i \mid \theta) + \sum_{i=1}^{n} \log p(x_i \mid z_i, \beta_{1:K})$$

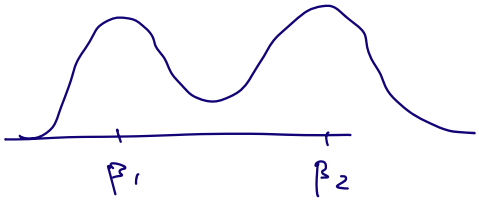
$\theta = 1/K$

n is huge

$$\sum_{i=1}^{n} \log p(x_i \mid z_i, \beta_{1:K}) = \sum_{i=1}^{n} \log f(x_i; \beta_{z_i})$$

$$\log f(x_i; \beta_k) = -\frac{1}{2\sigma^2} \sum_{j=1}^{d} (x_i^{(j)} - \beta_k^{(j)})^2 + \text{cnst.}$$

$$p(x \mid \beta_{1:K}) = \sum_{k=1}^{K} p(x \mid \beta_k)\, \theta_k$$



$$\log p(z_i \mid \theta) = -\log K$$

approximate posterior inference — Gibbs sampling
variational inference
MAP estimation

Gibbs sampler for a model

input: data
output: samples of latent variables from the posterior.

$$p(\theta, \beta_{1:K}, z_{1:n} \mid x_{1:n}) = \frac{p(\text{—})}{\underbrace{\int\int \sum_{z'_{1:n}} p(\theta', \beta'_{1:k}, z'_{1:n}, x_{1:n})\, d\beta'_{1:k}\, d\theta}_{p(x_{1:n})}}$$

- initialize θ, β, z to something
- iteratively sample from complete conditionals

$$\theta \sim p(\theta \mid \boldsymbol{\beta}, \mathbf{z}, \mathbf{x})$$
$$z_i \sim p(z_i \mid \theta, \boldsymbol{\beta}, \mathbf{z}_{-i}, \mathbf{x}) \quad i = 1..n$$
$$\beta_k \sim p(\beta_k \mid \theta, \boldsymbol{\beta}_{-k}, \mathbf{z}, \mathbf{x}) \quad k = 1..K$$

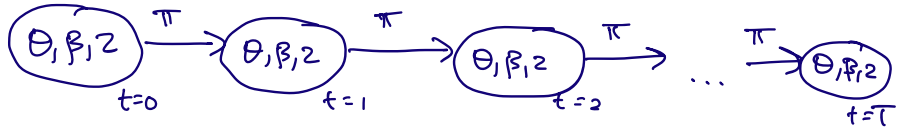
General model $p(\mathbf{h}, \mathbf{x})$ $\mathbf{h} = h_{1:d}$
Goal $p(\mathbf{h} \mid \mathbf{x})$
Gibbs $p(h_j \mid \mathbf{h}_{-j}, \mathbf{x})$ $j = 1..d$

Why does this recipe work?

MCMC - Markov chain Monte Carlo

$\{\theta, \beta, z\}_0 \sim \pi_0(\theta, \beta, z)$

$\{\theta, \beta, z\}_t \mid \{\theta, \beta, z\}_{t-1} \sim \pi(\theta, \beta, z \mid \{\theta, \beta, z\}_{t-1})$ ← transition distribution.



$\pi_t(\theta, \beta, z) \triangleq$ marginal at time t $\quad \pi_\infty(\theta, \beta, z) = p(\theta, \beta, z \mid x)$

Complete conditional of θ

$p(a,b) = p(a)\,p(b)$
$p(a|b) = p(a)$

$$p(\theta \mid \beta_{1:K}, z_{1:n}, x_{1:n}) \propto p(\theta, \beta_{1:K}, z_{1:n}, x_{1:n})$$

posterior $p(\theta \mid x_{1:n})$

$$\propto p(\theta) \prod_{i=1}^{n} p(z_i \mid \theta)$$

$p(\theta)$: Dirichlet; $p(z_i \mid \theta)$: Categorical

$\theta \perp\!\!\!\perp \{\beta_{1:K}, x_{1:n}\} \mid z_{1:n}$

$$n_k(\mathbf{z}) \triangleq \sum_{i=1}^{n} z_i^{(k)} = \text{Dirichlet}(\hat{\alpha}(\mathbf{z}))$$

$$\hat{\alpha}(\mathbf{z}) = \alpha + n(\mathbf{z}) \triangleq (n_1(\mathbf{z}), \ldots, n_K(\mathbf{z}))$$

Complete conditional of the assignments

$$p(z_i \mid \beta_{1:K}, \theta, \mathbf{z}_{-i}, x_{1:n}) \propto \text{joint distribution}$$

$$\propto p(z_i \mid \theta)\, p(x_i \mid z_i, \beta_{1:K})$$

$$= \theta_{z_i} f(x_i; \beta_{z_i})$$

$z_i \perp\!\!\!\perp \{\mathbf{z}_{-i}, x_{-i}\} \mid x_i, \beta_{1:K}, \theta$

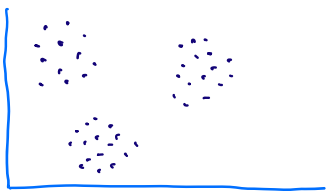
$$p(z_i = k \mid \text{---}) = \frac{\theta_k f(x_i; \beta_k)}{\sum_{k'=1}^{K} \theta_{k'} f(x_i; \beta_{k'})}$$

Complete conditional of the components

We know: how the data are partioned z_i

$p(\beta)$: Gaussian

$f(x_i; \beta)$: Gaussian w/ mean β (fixed variance)



$p(\beta_k \mid x_{1:n}, z_{1:n})$: Gaussian with "posterior" mean and variance from the EDA

$\Rightarrow$

- the practice of Gibbs sampling
- the history of " "
- the landscape of approximate inference.

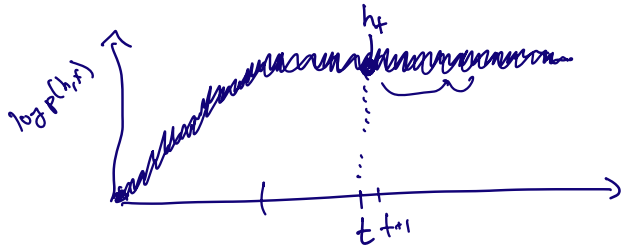
- Geman and Geman (1984): Gibbs for the Ising model
- Gelfand and Smith (1990): Gibbs for hierarchical Bayesian models

- burn-in, initialization, lag, convergence

$\pi_0 \to h_0 \xrightarrow{\pi} h_1 \xrightarrow{\pi} h_2 \to \ldots \xrightarrow{\pi} h_T$

π: transition
π_0: initial.
$\pi_T \triangleq$ marginal $\pi(h_T)$
$\lim_{t \to \infty} \pi_t(h_t) = p(h \mid x)$

monitor $\log p(h, x) \in$ deviance



$$p(h \mid x) \approx \sum_{b=1}^{B} \delta(h_b)$$

$h_b \sim p(h \mid x)$