Last week (Generative models for discrete)

- The Dirichlet model (for multiclass data)
- The Naive Bayes classifier

Today

- Generative models for continuous data
- Gaussian - Gaussian, Gaussian - Gamma models
- Bayesian regression
- Model selection & the Bayesian Occam's razor
- Gaussian Processes

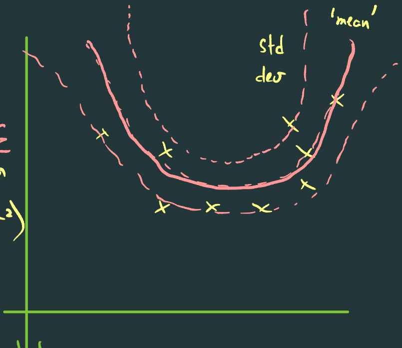
generative models for continuous data

Example – Regression (Bayesian)

Data: (x_i, t_i) , $i = 1, 2 \dots, n$

Model: $t = \underbrace{w_0 + w_1 x + w_2 x^2}_{\text{polynomial regression model}} + \varepsilon$

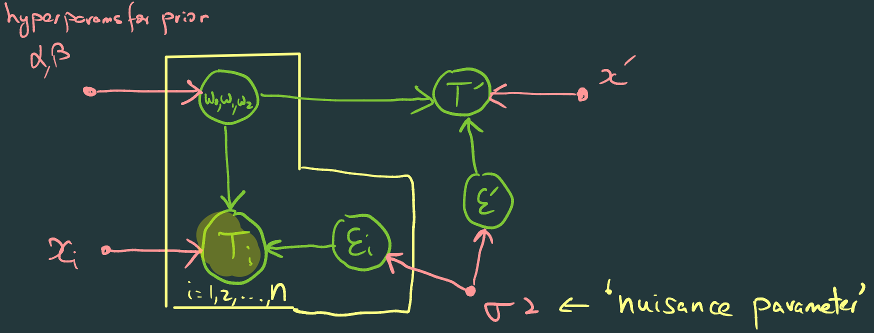
noise, $\mathcal{N}(0, \sigma^2)$



Q: find the 'best' degree 2 polynomial that 'approximates' data

Idea: Assume w_0, w_1, w_2 are random variables, learn from data via Bayesian update

Bayes Net for regression (for details, see Bishop Ch 8)



- $T' \perp\!\!\!\perp T_i \mid w$

- T' not independent of data $\{T_i\}$ if not conditioned on w

continuous data and Gaussian priors

- data $D = \{X_1, X_2, \ldots, X_n\} \in \mathbb{R}^n$, $X_i \in \mathbb{R}$ (1-dim data)
- model $\mathcal{M}$: X_i generated i.i.d. from $\mathcal{N}(\mu, \sigma^2)$ distribution

Gaussian prior

- $x \in \mathbb{R}$, parameters: $\Theta = (\mu, \sigma)$
- pdf: $\mathbb{P}[X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n] \propto \frac{1}{(\sigma^2)^{n/2}} \exp\left(\frac{-\sum_{i=1}^n (x_i - \mu)^2}{2\sigma^2}\right)$
- normalizing constant: $(2\pi)^{-n/2}$

conditioned on μ, σ^2

problem: functional form of μ and σ^2 are different

$\mu \in \mathbb{R}$, $\sigma^2 \in \mathbb{R}_+$

3 options:

1. μ unknown, σ^2 known   most important
2. σ^2 unknown, μ known
3. μ unknown, σ^2 unknown

notation: define precision $\tau = \frac{1}{\sigma^2}$

case 1: unknown μ

- data $D = \{X_1, X_2, \dots, X_n\} \in \mathbb{R}^n$ given constant, i.e., hyperparam
- model $\mathcal{M}$: X_i i.i.d. from $\mathcal{N}(\mu, 1/\tau)$, with unknown μ, known $\tau = 1/\sigma^2$

random variable

normal-normal model

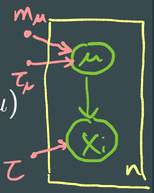
- likelihood: Gaussian likelihood fn

$$\mathcal{L}(\mu) = \boxed{p(D|\mu) \propto \tau^{n/2} \exp\left(-\tau \sum_{i=1}^{n} (x_i - \mu)^2 / 2\right)} = c_1 e^{-c_2(\mu - c_3)^2}$$

- prior parameter: $\Theta_0 = (m_\mu, 1/\tau_\mu)$ (mean, precision for μ)
- Gaussian prior for μ: $\mu \sim \mathcal{N}(m_\mu, \tau_\mu)$, where $\tau_\mu = 1/Var(\mu)$

hyperparams of prior

$$p(\mu | m_\mu, \tau_\mu) \propto \tau_\mu^{1/2} \exp\left(-\tau_\mu (\mu - m_\mu)^2 / 2\right)$$



normal-normal model: posterior

$$p(\mu \mid D) \propto \underbrace{p(\mu)}_{\text{prior}} \cdot \underbrace{\mathcal{L}_D(\mu)}_{\text{likelihood}}$$

- $A = \tau_\mu m_\mu + \tau \sum_{i=1}^{n} x_i$
- $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ = sample mean (MLE)
- $m_D = \dfrac{\tau_\mu m_\mu + n\tau \bar{x}}{\tau_\mu + n\tau}$
- $\tau_D = \tau_\mu + n\tau$

$$\propto \exp\left(-\frac{\tau_\mu (\mu - m_\mu)^2}{2}\right) \exp\left(\frac{-\tau}{2} \sum_{i=1}^{n} (x_i - \mu)^2\right)$$

want $e^{-c_1(\mu - c_2)^2}$

$$\propto \exp\left(\frac{-(\tau_\mu + n\tau)\mu^2 + 2A\mu + B}{2}\right)$$

$$\propto \exp\left(-\frac{\tau_D}{2}{}^2 (\mu - m_D)^2\right)$$

normal-normal model: posterior

$\Rightarrow$ posterior $\equiv$ $\mu \sim N(m_D, \tau_D)$

- posterior mean $m_D = \dfrac{\tau_\mu m_\mu + n\tau \bar{x}}{\tau_\mu + n\tau}$

$$= w_{prior}\, m_\mu + (1 - w_{prior}) \underbrace{\bar{x}}_{\text{MLE estimate}}$$

'shrinkage estimator'

- posterior precision $\tau_D = \tau_\mu + n\tau$

'precision on mean adds up under conditioning'

normal-normal model: posterior predictive distribution

- prior on $\mu \sim N(m_\mu, \tau_\mu)$, posterior given data $\mu_{|D} \sim N(m_D, \tau_D)$
- $X_{|\mu} \sim N(\mu, \tau) = \underbrace{\mu}_{\sim N(m_D, \tau_D)} + \frac{1}{\tau} N(0,1)$

$$= m_D + \frac{1}{\tau_D} N(0,1) + \frac{1}{\tau} N(0,1)$$

$$\sim N\Big(m_D, \underbrace{\frac{1}{\tau_D} + \frac{1}{\tau}}_{\sigma_D^2 + \sigma^2}\Big)$$

'posterior over $X \sim$ Gaussian, mean $\equiv$ convex comb of $m_\mu, \bar{x}$

variance $\equiv$ sum of σ_D^2, σ^2

ie. 'uncertainties add up for posterior prediction'

normal-normal model: posterior predictive distribution

- data $D = \{X_1, X_2, \dots, X_n\} \in \mathbb{R}^n$
- model $\mathcal{M}$: X_i i.i.d. from $\mathcal{N}(\mu, \tau)$, with unknown μ, known $\tau = 1/\sigma^2$
- thus we have $Z_1, Z_2 \sim N(0,1)$ iid

$$X_i = \mu + \sigma Z_1$$

$$\mu = m_\mu + \sigma_\mu Z_2 = \mu + \sigma Z_1 + \sigma_\mu Z_2$$

$$\Rightarrow \mathbb{E}[X_i] = \mathbb{E}[\mu] = m_D, \quad \mathrm{Var}(X_i) = \sigma^2 + \sigma_D^2$$

normal-normal model for unknown μ (see Bishop Ch 2, sec 3?)

- data $D = \{X_1, X_2, \ldots, X_n\} \in \mathbb{R}^n$ $\longrightarrow$ Suff stat $x_{MLE} = \bar{x} = \sum x_i / n$
- model $\mathcal{M}$: X_i i.i.d. from $\mathcal{N}(\mu, \tau)$, with unknown μ, known $\tau = 1/\sigma^2$

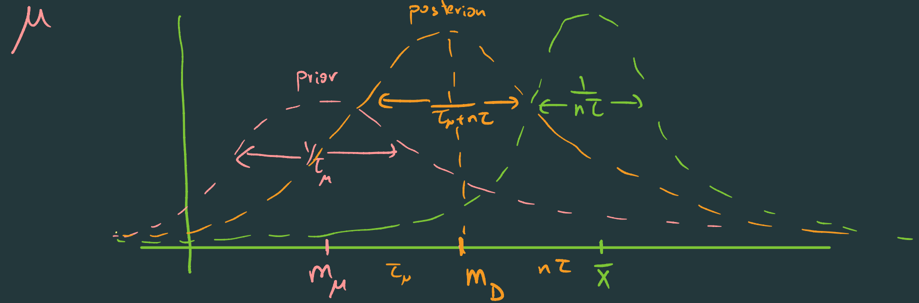
normal-normal model

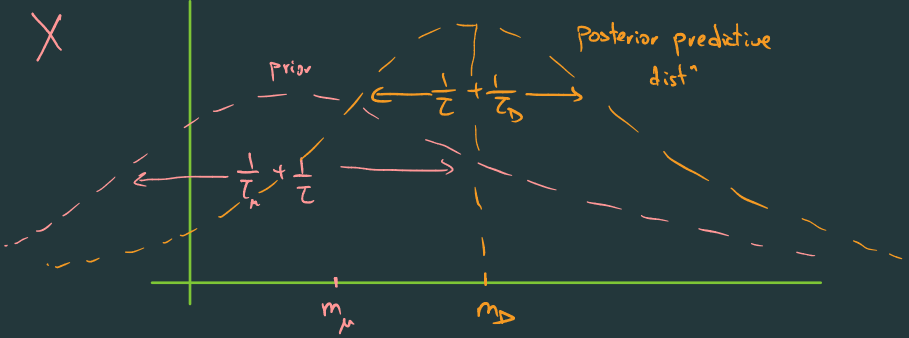
- likelihood: $p(D|\mu) \propto \exp\left(-\tau \sum_{i=1}^{n} (x_i - \mu)^2 / 2\right)$
- prior: $\mu \sim \mathcal{N}(m_\mu, 1/\tau_\mu) \propto \exp\left(-\tau_\mu (\mu - m_\mu)^2 / 2\right)$
- posterior: let $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$, $m_D = \frac{n\tau \cdot \bar{x} + \tau_\mu \cdot m_\mu}{n\tau + \tau_\mu}$ and $\underbrace{\tau_D = n\tau + \tau_\mu}_{\text{precisions add for inference}}$

$$p(\mu|D) \sim \mathcal{N}\left(m_D, 1/\tau_D\right)$$

- posterior predictive distribution:

$$p(x|D) \sim \mathcal{N}\left(m_D, \underbrace{1/\tau + 1/\tau_D}_{\text{variances add for prediction}}\right)$$


μ
posterior
prior
$\frac{1}{\tau_\mu + n\tau}$
$\frac{1}{n\tau}$
$1/\tau_\mu$
m_μ
τ_μ
m_D
$n\tau$
$\bar{X}$

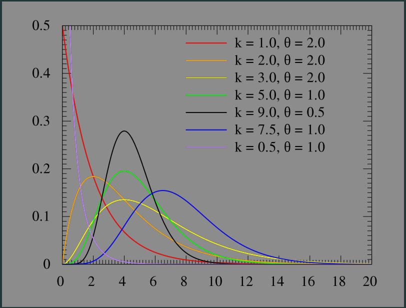

X
prior
Posterior predictive distⁿ
$\frac{1}{\tau} + \frac{1}{\tau_D}$
$\frac{1}{\tau_\mu} + \frac{1}{\tau}$
m_μ
m_D

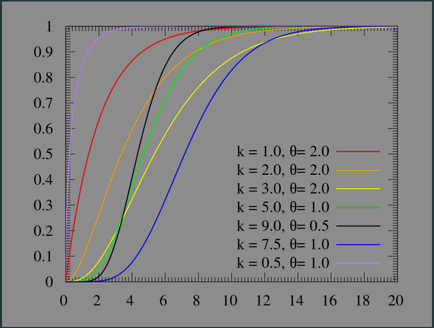
the gamma distribution

gamma distribution

- $x \in (0, \infty)$, parameters: $\Theta = (\alpha, \beta) \in \mathbb{R}^+$ ('shape,rate')
- pdf of $Gamma(\alpha, \beta)$: $p(x) \propto x^{\alpha-1} e^{-\beta x}$
- normalizing constant: $\frac{1}{Z(\alpha,\beta)} = \frac{\beta^\alpha}{\Gamma(\alpha)}$

Eg - If $X_i \sim Exp(\lambda)$ iid, then $X_1 + X_2 \sim$ Gamma with mean $2/\lambda$





case 2: unknown σ

- data $D = \{X_1, X_2, \ldots, X_n\} \in \mathbb{R}^n$
- model $\mathcal{M}$: X_i i.i.d. from $\mathcal{N}(\mu, 1/\tau)$, with unknown $\tau = 1/\sigma$, known μ

normal-gamma model

- likelihood:

$$p(D|\theta) \propto \tau^{n/2} \exp\left(-\tau \sum_{i=1}^{n} (x_i - \mu)^2/2\right)$$

- prior parameters: $\Theta_0 = (\alpha, \beta)$
- gamma prior for τ: $\tau \sim Gamma(\alpha, \beta)$

$$p(\tau|\alpha, \beta) \propto \tau^{\alpha-1} e^{-\beta\tau}$$

normal-gamma model: posterior

$$p(\tau|D) \propto \underbrace{\tau^{\alpha-1} e^{-\beta\tau}}_{\text{prior}} \underbrace{\tau^{n/2} e^{-\frac{\tau}{2}\sum(x_i-\mu)^2}}_{\text{Likelihood}}$$

$$\propto \tau^{\alpha+\frac{n}{2}-1} e^{-\tau\left(\beta+\frac{\sum(x_i-\mu)^2}{2}\right)}$$

$$= \text{Gamma}\left(\alpha+\frac{n}{2}, \beta+\frac{\sum(x_i-\mu)^2}{2}\right)$$

normal-gamma model: posterior predictive distribution

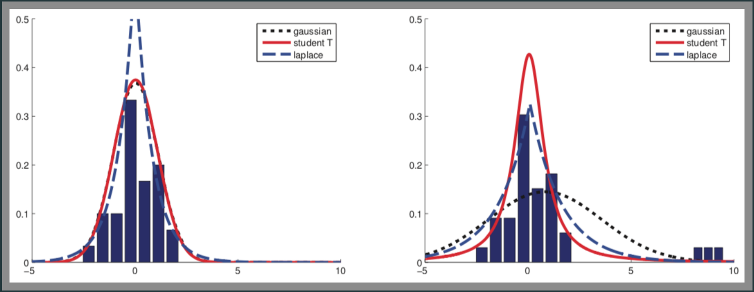
$$p(x \mid \alpha, \beta, D) = \int_0^\infty \underbrace{p(\tau \mid \alpha, \beta, D)}_{\text{gamma}} \cdot \underbrace{p(x \mid \tau)}_{\text{Gaussian}} \, d\tau$$

$$= \underbrace{\frac{1}{Z} \; \frac{1}{\left(1 + \frac{(x-\mu)^2}{2\beta}\right)^{\alpha_D + 1/2}}}_{\text{student's t-dist}}$$

the Student-t distribution ('naturally robust' distribution)

Student-t distribution

- $x \in \mathbb{R}$, parameter: $\mu \in \mathbb{R}, \nu > 0$ (mean,'degrees of freedom')
- pdf of student-t(μ, ν): $p(x) \propto \left(1 + \frac{(x-\mu)^2}{\nu}\right)^{\frac{\nu+1}{2}}$
- normalizing constant: $\frac{1}{Z(\mu,\nu)} = \frac{\Gamma(\nu+1)/2)}{\sqrt{\nu\pi}\Gamma(\nu/2)}$



robustness of student-t to outliers

normal-gamma model for unknown τ

- data $D = \{X_1, X_2, \ldots, X_n\} \in \mathbb{R}^n$
- model $\mathcal{M}$: X_i i.i.d. from $\mathcal{N}(\mu, \sigma^2)$, with unknown $\tau = 1/\sigma^2$, known μ

normal-gamma model

- likelihood: $p(D|\theta) \propto \exp\left(-\tau \sum_{i=1}^{n}(x_i - \mu)^2/2\right)$
- prior for τ: $\tau \sim gamma(\alpha, \beta)$
- posterior: let $\alpha_D = \alpha + \frac{n}{2}$ and $\beta_D = \beta + \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2$

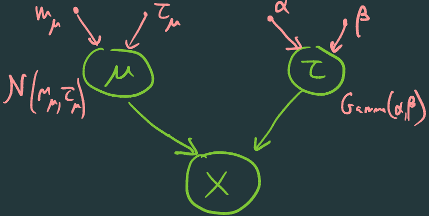
$$p(\tau|D) \sim gamma\left(\alpha_D, \beta_D\right)$$

- posterior predictive distribution:

$$p(x|D) \sim \text{student-t}$$

case 3: unknown μ and σ^2

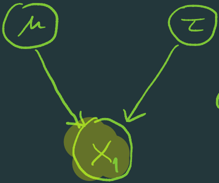
Idea 1



Assume μ and τ are independant

Now given data X_1, what happens to μ, τ ?

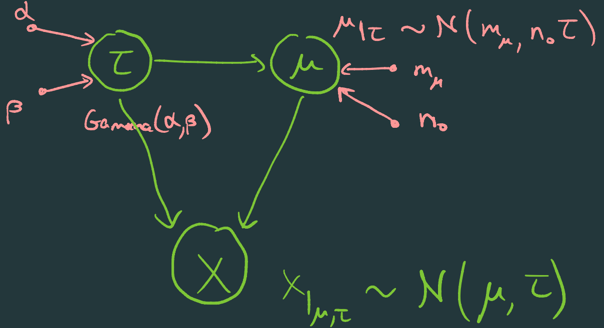
want posterior to be



this is not true!

Conditioned on X_1, μ ad τ are no longer indep ('explaining away')

case 3: unknown μ and σ^2



Bayesian update for this is known in closed form ('correct' conjugate prior)