Introduction to Machine Learning - CS725
Instructor: Prof. Ganesh Ramakrishnan
Lecture 6 - Bayesian Inference for (Multivariate) Gaussian and Bayesian Linear Regression

Recap: Bayesian Inference for Bernoulli

Let $\mathcal{D} \mid H$ follow a distribution $Ber(p)$ (p is probability of heads) and p follow a distribution $Beta(p; \alpha, \beta) \sim \frac{p^{(\alpha-1)}(1-p)^{(\beta-1)}}{B(\alpha,\beta)}$,

1. *The Maximum Likelihood Estimate:*
   $$\hat{p} = \underset{p}{\operatorname{argmax}} \; {}^{n}C_h p^h (1-p)^{n-h} = \frac{h}{n}$$
2. *The Posterior Distribution:*
   $$\Pr(p \mid \mathcal{D}) = Beta(p; \alpha + h, \beta + n - h)$$
3. *The Maximum a-Posterior (MAP) Estimate:* The mode of the posterior distribution
   $$\tilde{p} = \underset{H}{\operatorname{argmax}} \Pr(H \mid \mathcal{D}) = \underset{p}{\operatorname{argmax}} \Pr(p \mid \mathcal{D})$$
   $$= \operatorname{argmax} Beta(p; \alpha + h, \beta + n - h) = \frac{\alpha + h - 1}{\alpha + \beta + n - 2}$$

Recap: Conjugate Prior for (univariate) Gaussian

- Let $\Pr(X) \sim \mathcal{N}(\mu, \sigma^2)$ and let the data $\mathcal{D} = x_1 \ldots x_m$
- $\mu_{MLE} = \frac{1}{m}\sum_{i=1}^{m} x_i$ and $\sigma^2_{MLE} = \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_{MLE})^2$
- Suppose you are told that the conjugate prior for the (univariate) normally distributed random variable X in the case that σ^2 is not a random variable is $\Pr(\mu) = \mathcal{N}(\mu_0, \sigma_0^2)$. Then the **posterior** is?
- Answer: $\Pr(\mu | x_1 \ldots x_m) = \mathcal{N}(\mu_m, \sigma_m^2)$ such that $\mu_m = \ldots\ldots$ and $\frac{1}{\sigma_m^2} = \ldots\ldots$

  $\because$ μ_{MLE} is linear, $\mu_m = \theta_1 \mu_{MLE} + \theta_2 \mu_0$ is expected
- Helpful tip: Product of Gaussians is always a Gaussian

$$\Pr(\mu) = \frac{1}{\sqrt{2\pi\sigma_0^2}} exp\left(\frac{-(\mu-\mu_0)^2}{2\sigma_0^2}\right)$$

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

$\prod_i \Pr(x_i|\mu, \sigma^2)$

$$\Pr(\mathcal{D}|\mu) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^m exp\left(\frac{-1}{2\sigma^2}\sum_{i=1}^{m}(x_i-\mu)^2\right)$$

[$\because$ $x_1 . x_m$ are iid]

$$\Pr(\mu|\mathcal{D}) \propto \Pr(\mathcal{D}|\mu)\Pr(\mu) =$$

$$\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^m \frac{1}{\sqrt{2\pi\sigma_0^2}} exp\left(\frac{-1}{2\sigma^2}\sum_{i=1}^{m}(x_i-\mu)^2 - \frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right) \propto$$

$$exp\left(\frac{-1}{2\sigma^2}\sum_{i=1}^{m}(x_i-\mu)^2 - \frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right) = exp\left(\frac{-1}{2\sigma_m^2}(\mu-\mu_m)^2\right)$$

Recap: Detailed derivation (contd.)

Our reference equality:

$$exp\left(\frac{-1}{2\sigma^2}\sum_{i=1}^{m}(x_i-\mu)^2-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right)=exp\left(\frac{-1}{2\sigma_m^2}(\mu-\mu_m)^2\right),$$

Matching coefficients of μ^2, we get

$$\left(\sum_{i=1}^{m}-\frac{1}{2\sigma^2}\right)-\frac{1}{2\sigma_0^2}=\frac{-1}{2\sigma_m^2}\quad\Big\}\quad\frac{m}{\sigma^2}+\frac{1}{\sigma_0^2}=\frac{1}{\sigma_m^2}$$

As $m\to\infty$, $\sigma_m^2\propto\frac{1}{m}$

Recap: Detailed derivation (contd.)

Our reference equality:

$$exp\left(\frac{-1}{2\sigma^2}\sum_{i=1}^{m}(x_i-\mu)^2-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right)=exp\left(\frac{-1}{2\sigma_m^2}(\mu-\mu_m)^2\right),$$

Matching coefficients of μ^2, we get

$\frac{-\mu^2}{2\sigma_m^2}=\frac{-\mu^2}{2}(\frac{m}{\sigma^2}+\frac{1}{\sigma_0^2})\Rightarrow$

Recap: Detailed derivation (contd.)

Both μ & σ^2 are either random vars or known!

Our reference equality:

$$exp\left(\frac{-1}{2\sigma^2}\sum_{i=1}^{m}(x_i-\mu)^2-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right)=exp\left(\frac{-1}{2\sigma_m^2}(\mu-\mu_m)^2\right),$$

Matching coefficients of μ^2, we get

$$\frac{-\mu^2}{2\sigma_m^2}=\frac{-\mu^2}{2}\left(\frac{m}{\sigma^2}+\frac{1}{\sigma_0^2}\right)\Rightarrow\frac{1}{\sigma_m^2}=\frac{1}{\sigma_0^2}+\frac{m}{\sigma^2}$$

$\to$ can we substitute & eliminate?

Ans: Not in Bayesian estimation. If σ^2 is NOT a r.v. it must be known

Matching coefficients of μ, we get

$$\frac{-1}{2\sigma^2}\left(\sum_{i=1}^{m}-2x_i\right)-\frac{(-2\mu_0)}{2\sigma_0^2}=-\frac{(-2\mu_m)}{2\sigma_m^2}$$

$$\Rightarrow\left(\sum_{i=1}^{m}x_i\right)/\sigma^2+\mu_0/\sigma_0^2=\mu_m/\sigma_m^2$$

mMLE

Recap: Detailed derivation (contd.)

Our reference equality:

$$exp\left(\frac{-1}{2\sigma^2}\sum_{i=1}^{m}(x_i-\mu)^2-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right)=exp\left(\frac{-1}{2\sigma_m^2}(\mu-\mu_m)^2\right),$$

Matching coefficients of μ^2, we get

$$\frac{-\mu^2}{2\sigma_m^2}=\frac{-\mu^2}{2}\left(\frac{m}{\sigma^2}+\frac{1}{\sigma_0^2}\right)\Rightarrow\frac{1}{\sigma_m^2}=\frac{1}{\sigma_0^2}+\frac{m}{\sigma^2}$$

Matching coefficients of μ, we get

$$\frac{2\mu\mu_m}{2\sigma_m^2}=\mu\left(\frac{2\sum_{i=1}^{m}x_i}{2\sigma^2}+\frac{2\mu_0}{2\sigma_0^2}\right)\Rightarrow$$

Recap: Detailed derivation (contd.)

Our reference equality:

$$exp\left(\frac{-1}{2\sigma^2}\sum_{i=1}^{m}(x_i-\mu)^2-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right)=exp\left(\frac{-1}{2\sigma_m^2}(\mu-\mu_m)^2\right),$$

Matching coefficients of μ^2, we get

$$\frac{-\mu^2}{2\sigma_m^2}=\frac{-\mu^2}{2}\left(\frac{m}{\sigma^2}+\frac{1}{\sigma_0^2}\right)\Rightarrow\frac{1}{\sigma_m^2}=\frac{1}{\sigma_0^2}+\frac{m}{\sigma^2}$$

Matching coefficients of μ, we get

$$\frac{2\mu\mu_m}{2\sigma_m^2}=\mu\left(\frac{2\sum_{i=1}^{m}x_i}{2\sigma^2}+\frac{2\mu_0}{2\sigma_0^2}\right)\Rightarrow\mu_m=\sigma_m^2\left(\frac{\sum_{i=1}^{m}x_i}{\sigma^2}+\frac{\mu_0}{\sigma_0^2}\right)\text{ or}$$

$$\mu_m=\sigma_m^2\left(\frac{m\hat{\mu}_{ML}}{\sigma^2}+\frac{\mu_0}{\sigma_0^2}\right)\Rightarrow$$

Recap: Detailed derivation (contd.)

Our reference equality:

$$exp\left(\frac{-1}{2\sigma^2}\sum_{i=1}^{m}(x_i-\mu)^2 - \frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right) = exp\left(\frac{-1}{2\sigma_m^2}(\mu-\mu_m)^2\right),$$

Matching coefficients of μ^2, we get

$$\frac{-\mu^2}{2\sigma_m^2} = \frac{-\mu^2}{2}\left(\frac{m}{\sigma^2}+\frac{1}{\sigma_0^2}\right) \Rightarrow \frac{1}{\sigma_m^2} = \frac{1}{\sigma_0^2}+\frac{m}{\sigma^2}$$

$\theta_1 = \frac{m\sigma_0^2}{m\sigma_0^2+\sigma^2}$

$\theta_2 = \frac{\sigma^2}{m\sigma_0^2+\sigma^2}$

Matching coefficients of μ, we get

$$\frac{2\mu\mu_m}{2\sigma_m^2} = \mu\left(\frac{2\sum_{i=1}^{m}x_i}{2\sigma^2}+\frac{2\mu_0}{2\sigma_0^2}\right) \Rightarrow \mu_m = \sigma_m^2\left(\frac{\sum_{i=1}^{m}x_i}{\sigma^2}+\frac{\mu_0}{\sigma_0^2}\right) \text{ or}$$

$$\mu_m = \sigma_m^2\left(\frac{m\hat{\mu}_{ML}}{\sigma^2}+\frac{\mu_0}{\sigma_0^2}\right) \Rightarrow \boxed{\mu_m = \left(\frac{\sigma^2}{m\sigma_0^2+\sigma^2}\mu_0\right)+\left(\frac{m\sigma_0^2}{m\sigma_0^2+\sigma^2}\hat{\mu}_{ML}\right)}$$

Summary: Conjugate Prior for (univariate) Gaussian

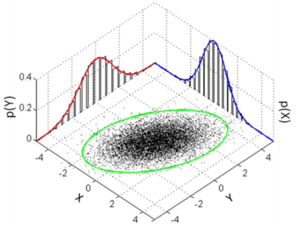
- Let $\Pr(X) \sim \mathcal{N}(\mu, \sigma^2)$ and let the data $\mathcal{D} = x_1 \ldots x_m$
- $\mu_{MLE} = \frac{1}{m}\sum_{i=1}^{m} x_i$ and $\sigma^2_{MLE} = \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_{MLE})^2$
- Suppose σ^2 is not a random variable and $\Pr(\mu) = \mathcal{N}(\mu_0, \sigma_0^2)$. Then:
- $\Pr(\mu | x_1 \ldots x_m) = \mathcal{N}(\mu_m, \sigma_m^2)$ such that

Precision increases as m increases

- $\mu_m = \left(\frac{\sigma^2}{m\sigma_0^2 + \sigma^2}\mu_0\right) + \left(\frac{m\sigma_0^2}{m\sigma_0^2 + \sigma^2}\hat{\mu}_{ML}\right)$ and $\frac{1}{\sigma_m^2} = \frac{1}{\sigma_0^2} + \frac{m}{\sigma^2}$
where $1/\sigma_0^2$ can be attributed to uncertainty in μ and m/σ^2 can be attributed to noise in observation

Multivariate Normal Distribution and MLE estimate

1. The multivariate Gaussian (Normal) Distribution is: $\mathcal{N}(\mathbf{x};\mu,\Sigma) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)}$ when $\Sigma \in \Re^{n\times n}$ is positive-definite and $\mu \in \Re^n$



$x = [x_1 \cdots x_n]$

let $x_1 .. x_n \sim \mathcal{N}(\cdot)$ individually

Also linear combinations of subsets of $\{x_1 .. x_n\} \sim \mathcal{N}(\cdot)$

xi's are neither assumed to be independent nor have identical distribution

Multivariate Normal Distribution and MLE estimate

Can be ignored for the time being

1. The multivariate Gaussian (Normal) Distribution is: $\mathcal{N}(\mathbf{x};\mu,\Sigma) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)}$ when $\Sigma \in \Re^{n\times n}$ is positive-definite and $\mu \in \Re^n$
2. $\mu_{MLE} = \boxed{\frac{1}{m}\sum_{i=1}^{m}\mathbf{x}_i} \sim \frac{1}{m}\sum_{i=1}^{m}\phi(\mathbf{x}_i)$ and

   If you treated x_i's as abstract objects represented in vector space of feature vectors $\phi(.)$

   $\underbrace{\Sigma_{MLE}}_{\text{Avg of several rank 1 approx matrices}} \sim \frac{1}{m}\sum_{i=1}^{m}\underbrace{(\phi(\mathbf{x}_i)-\mu_{MLE})}_{[v]}(\phi(\mathbf{x}_i)-\mu_{MLE})^T$

   $\begin{bmatrix} x_i-\mu_{MLE} \end{bmatrix}\underbrace{\begin{bmatrix} x_i-\mu_{MLE} \end{bmatrix}}_{[\alpha_1\ \alpha_2\ \ldots\ \alpha_n]} = \frac{1}{m}\sum \underbrace{\begin{bmatrix} \alpha_1 v & \lambda_2 v & \ldots & \alpha_n v \end{bmatrix}}_{\text{Rank 1 matrix}}$

Summary for MAP estimation with Normal Distribution

Expect univariate to be special case of multivariate

- Univariate: With $\mu \sim \mathcal{N}(\mu_0, {\sigma^2}_0)$ and $x \sim \mathcal{N}(\mu, \sigma^2)$, $p(x|D) \sim \mathcal{N}(\mu_m, {\sigma_m}^2)$

$$\Sigma_m^{-1} \leftarrow \frac{1}{\sigma_m^2} = \frac{m}{\sigma^2} + \frac{1}{\sigma_0^2} \rightarrow \Sigma_0^{-1}$$

for 1 d case $\frac{1}{\sigma^2} = \Sigma^{-1}$

$$\Sigma_m^{-1}\mu_m \quad \frac{\mu_m}{\sigma_m^2} = \frac{m}{\sigma^2}\hat{\mu}_{mle} + \frac{\mu_0}{\sigma_0^2}$$

- Multivariate: By **extrapolation** (Bayesian setting for fixed Σ)

$$\mathbf{x} \sim \mathcal{N}(\mu, \Sigma), \ \mu \sim \mathcal{N}(\mu_0, \Sigma_0) \ \& \ p(\mathbf{x}|D) \sim \mathcal{N}(\mu_m, \Sigma_m)$$

$\mu_m^T \Sigma_m^{-1} = \Sigma_m^{-1} \mu_m$

$\because \Sigma_m$ is symmetric psd

$$\Sigma_m^{-1} = m\Sigma^{-1} + \Sigma_0^{-1}$$

(can prove that Σ_m^{-1} is invertible)

$$\Sigma_m^{-1}\mu_m = m\Sigma^{-1}\mu_{MLE} + \Sigma_0^{-1}\mu_0$$

Summary for MAP estimation with Normal Distribution

- Univariate: With $\mu \sim \mathcal{N}(\mu_0, {\sigma^2}_0)$ and $x \sim \mathcal{N}(\mu, \sigma^2)$,$p(x|D) \sim \mathcal{N}(\mu_m, {\sigma_m}^2)$

$$\frac{1}{\sigma_m^2} = \frac{m}{\sigma^2} + \frac{1}{\sigma_0^2}$$
$$\frac{\mu_m}{\sigma_m^2} = \frac{m}{\sigma^2}\hat{\mu}_{mle} + \frac{\mu_0}{\sigma_0^2}$$

- Multivariate: By **extrapolation** (Bayesian setting for fixed Σ)

$$\mathbf{x} \sim \mathcal{N}(\mu, \Sigma),\ \ \mu \sim \mathcal{N}(\mu_0, \Sigma_0)\ \&\ p(\mathbf{x}|D) \sim \mathcal{N}(\mu_m, \Sigma_m)$$

$$\Sigma_m^{-1} = m\Sigma^{-1} + \Sigma_0^{-1}$$
$$\Sigma_m^{-1}\mu_m = m\Sigma^{-1}\hat{\mu}_{mle} + \Sigma_0^{-1}\mu_0$$

Different Estimators

	Point?	$p(x\|D)$
MLE ✓	$\hat{\theta}_{MLE} = \text{argmax}_\theta \, LL(D\|\theta)$	$p(x\|\theta_{MLE})$
Bayes Estimator	$\hat{\theta}_B = E_{p(\theta\|D)} E[\theta]$	$p(x\|\theta_B)$
MAP ✓	$\hat{\theta}_{MAP} = \text{argmax}_\theta \, p(\theta\|D)$	$p(x\|\theta_{MAP})$
Pure Bayesian	 $\theta_{prior} \rightarrow D \rightarrow \theta_{post}$; $\theta_{post} \rightarrow$ query pt x ; $\Pr(x\|D) \quad \theta \leftarrow$ query pt x	$p(\theta\|D) = \dfrac{p(D\|\theta)p(\theta)}{\int_m p(D\|\theta)p(\theta)d\theta}$ $p(D\|\theta) = \prod_{i=1} p(x_i\|\theta)$ $p(x\|D) = \int_\theta p(x\|\theta)p(\theta\|D)$

Back to Linear Regression: Why Bayesian?

- The Bayesian interpretation of probabilistic estimation is a logical extension that enables reasoning with uncertainty **but in the light of some background belief**
- **Bayesian linear regression**: A Bayesian alternative to **Maximum Likelihood** least squares regression to address **overfitting**
- Continue with Normally distributed errors
- Model the $\mathbf{w}$ using a prior distribution and use the posterior over $\mathbf{w}$ as the result
- Intuitive Prior: Components of $\mathbf{w}$ should not become too large!

Back to Linear Regression: Prior Distribution for $\mathbf{w}$

$$y = \mathbf{w}^T \phi(\mathbf{x}) + \varepsilon$$
$$\varepsilon \sim \mathcal{N}(0, \sigma^2)$$

- Maximum (log)-likelihood estimate is $\hat{\mathbf{w}}_{MLE} = (\Phi^T \Phi)^{-1} \Phi^T y$
- We can use a Prior distribution on $\mathbf{w}$ to avoid over-fitting

$w_i \sim N(0, \frac{1}{\lambda})$ OR $N(0, \sigma^2)$

99% $w_i \in \left[-\frac{3}{\sqrt{\lambda}}, +\frac{3}{\sqrt{\lambda}}\right]$

$w \in N(0, \frac{1}{\lambda} I)$

H/W: Find posterior on w

Back to Linear Regression: Prior Distribution for $\mathbf{w}$

$$y = \mathbf{w}^T \phi(\mathbf{x}) + \varepsilon$$
$$\varepsilon \sim \mathcal{N}(0, \sigma^2)$$

- Maximum (log)-likelihood estimate is $\hat{\mathbf{w}}_{MLE} = (\Phi^T \Phi)^{-1} \Phi^T y$
- We can use a Prior distribution on $\mathbf{w}$ to avoid over-fitting
$$w_i \sim \mathcal{N}(0, \tfrac{1}{\lambda})$$
(that is, each component w_i is approximately bounded within $\pm\frac{3}{\sqrt{\lambda}}$ by the $3 - \sigma$ rule). λ is also called the precision of the Gaussian
- Q: Bayesian Estimation?