Introduction to Machine Learning - CS725
Instructor: Prof. Ganesh Ramakrishnan
Lecture 6 - Linear Regression - Bayesian Inference and Regularization

Building on questions on Least Squares Linear Regression

1. Is there a probabilistic interpretation?
   - Gaussian Error, Maximum Likelihood Estimate
2. Addressing overfitting
   - Bayesian and Maximum Aposteriori Estimates, Regularization
3. How to minimize the resultant and more complex error functions?
   - Level Curves and Surfaces, Gradient Vector, Directional Derivative, Gradient Descent Algorithm, Convexity, Necessary and Sufficient Conditions for Optimality

Recap: Bayesian Inference with Coin Tossing

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

$E_{\text{Posterior}}[p] = \int p \, Beta(p; \alpha+h, \beta+n-h) = \frac{\alpha+h}{\alpha+\beta+n}$

Bayes Estimate

Intuition for Bayesian Linear Regression

- 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

  A 3rd probabilistic form

- Continue with Normally distributed errors
- Model the **w** using a prior distribution and use the posterior over **w** as the result
- Intuitive Prior: Components of **w** should not become too large!

2 forms of regularized Ridge regression:

1. $w^* = \arg\min_{w} \|\phi w - y\|_2^2$ s.t. $\|w\|_2^2 \leq \theta$
2. $w^* = \arg\min_{w} \|\phi w - y\|_2^2 + \lambda \|w\|_2^2$

$\phi = m \times n$ matrix

Prior Distribution for $\mathbf{w}$ for Linear Regression

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

- We saw that when we try to maximize log-likelihood we end up with $\hat{\mathbf{w}}_{MLE} = (\Phi^T \Phi)^{-1} \Phi^T y$
- We can use a Prior distribution on $\mathbf{w}$ to avoid over-fitting

$$w_i \sim \mathcal{N}(0, \tfrac{1}{\lambda})$$

in absence of D,
$E(w_i) = 0$

Each component w_i is approximately bounded within $\pm \frac{3}{\sqrt{\lambda}}$. λ is also called the precision of the Gaussian

increasing λ

- Q1: How do deal with Bayesian Estimation for Gaussian distribution? $p(w_i | D)$?

Conjugate Prior for (univariate) Gaussian

- We will temporarily generalize the discussion with x taking the place of ε and μ taking the place of w_i

Conjugate Prior for (univariate) Gaussian

- We will temporarily generalize the discussion with x taking the place of ε and μ taking the place of w_i
- 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?

Conjugate Prior for (univariate) Gaussian

- We will temporarily generalize the discussion with x taking the place of ε and μ taking the place of w_i
- 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 $\dfrac{1}{\sigma_m^2} = \ldots..$
- Helpful tip: Product of Gaussians is always a Gaussian

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

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

$$\underbrace{\Pr(x_1 \ldots x_m \mid \mu)}_{\text{under iid assumption}} = \prod_{i=1}^{m} \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)$$

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

$$\Pr(\mu \mid x_1 \ldots x_m) = \frac{\Pr(x_1 \ldots x_m \mid \mu)\,\Pr(\mu)}{\underbrace{\Pr(x_1 \ldots x_m)}_{\text{A normalizing factor}}} \propto \Pr(x_1 \ldots x_m \mid \mu)\,\Pr(\mu)$$

& we will bundle all terms independent of μ into normalization

Detailed derivation

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

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

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

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

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

Further ignore terms independent of μ

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

My leap of faith: $\Pr(\mu|x_1 \ldots x_m) = \mathcal{N}(\mu_m, \sigma_m^2)$

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

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

Precision grows as data observed increases.

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

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

Detailed derivation (contd.)

Our reference equality:

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

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

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

$$\frac{\mu_m}{\sigma_m^2}=\frac{\mu_0}{\sigma_0^2}+\frac{\sum x_i}{\sigma^2}=\frac{\mu_0}{\sigma_0^2}+\frac{\mu_{MLE} * m}{\sigma^2}$$

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

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

Matching coefficients of μ, we get

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

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

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

Matching coefficients of μ, we get

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

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

$\hat{\mu}_{MLE} * m = \sum_{i=1}^{m} x_i$

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

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

$$\mu_m=\sigma_m^2\left(\frac{m\hat{\mu}_{ML}}{\sigma^2}+\frac{\mu_0}{\sigma_0^2}\right)\Rightarrow\mu_m=\Big(\underbrace{\frac{\sigma^2}{m\sigma_0^2+\sigma^2}}_{\text{wt}\to 0 \text{ as } m\to\infty}\mu_0\Big)+\Big(\underbrace{\frac{m\sigma_0^2}{m\sigma_0^2+\sigma^2}}_{\text{wt}\to 1 \text{ as } m\to\infty}\hat{\mu}_{ML}\Big)$$

Data starts dominating

How abt: $w\in\mathbb{R}^n$ st $w_i \in \mathcal{N}(0,\frac{1}{\lambda})$

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 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

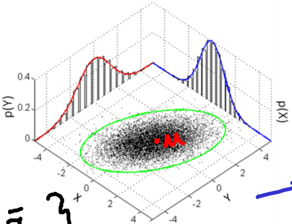
Summary: Conjugate Prior for (univariate) Gaussian

- Let $\Pr(X) \sim \mathcal{N}(\mu, \sigma^2)$ and let the data $\mathcal{D} = x_1 ... x_m$
- $\mu_{MLE} = \frac{1}{m}\sum_{i=1}^{m} x_i$ 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: $\Pr(\mu | x_1 ... x_m) = \mathcal{N}(\mu_m, \sigma_m^2)$ such that
- $\mu_m = (\frac{\sigma^2}{m\sigma_0^2 + \sigma^2}\mu_0) + (\frac{m\sigma_0^2}{m\sigma_0^2 + \sigma^2}\hat{\mu}_{ML})$
- $\frac{1}{\sigma_m^2} = \frac{1}{\sigma_0^2} + \frac{m}{\sigma^2}$

Multivariate Normal Distribution and MLE estimate

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

$P_r(x_1) = \int_{x_2..x_n} N(x, \mu, \Sigma)\, dx_2 .. dx_n = N(x_1, \mu_1, \sigma_1^2)$



plane of $x \in \Re^2$ points scattered around μ as per 3σ rule

$D = \{\bar{x}_1 .. \bar{x}_m\}$

Weka workbench for ML: Visualization in pairs of (x_i, x_j)

2. $\mu_{MLE} = \frac{1}{m}\sum_{i=1}^{m} \mathbf{x}_i \sim \frac{1}{m}\sum_{i=1}^{m} \phi(\mathbf{x}_i)$ and

$$\Sigma_{MLE} \sim \frac{1}{m}\sum_{i=1}^{m} (\phi(\mathbf{x}_i) - \mu_{MLE})(\phi(\mathbf{x}_i) - \mu_{MLE})^T = [\vdots][\cdots]$$

$$(\Sigma_{MLE})_{kj} = \frac{1}{m}\sum_{i=1}^{m} (x_{ik} - (\mu_{MLE})_k)(x_{ij} - (\mu_{MLE})_j)$$

Summary for MAP estimation with Normal Distribution

Univariate Case

- Summary: With $\mu \sim \mathcal{N}(\mu_0, {\sigma^2}_0)$ and $x \sim \mathcal{N}(\mu, \sigma^2)$

$$\frac{1}{\sigma_m^2} = \frac{m}{\sigma^2} + \frac{1}{\sigma_0^2}$$

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

such that $p(x|D) \sim \mathcal{N}(\mu_m, {\sigma_m}^2)$. Here m/σ^2 is due to noise in observation while $1/\sigma_0^2$ is due to uncertainity in μ

- For the Bayesian setting for the multivariate case with fixed Σ

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

$\Sigma \in \mathbb{R}^{n \times n}$ $\quad \mu \in \mathbb{R}^n$ $\quad \mu_0 \in \mathbb{R}^n$ $\quad \Sigma_0 \in \mathbb{R}^{n \times n}$

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

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

If $n=1$

$$\Sigma_m = (\sigma_m)^2$$

$$\Sigma_m^{-1} = \left(\frac{1}{\sigma_m}\right)^2$$

Summary for MAP estimation with Normal Distribution

- Summary: With $\mu \sim \mathcal{N}(\mu_0, {\sigma^2}_0)$ and $x \sim \mathcal{N}(\mu, \sigma^2)$

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

  such that $p(x|D) \sim \mathcal{N}(\mu_m, {\sigma_m}^2)$. Here m/σ^2 is due to noise in observation while $1/\sigma_0^2$ is due to uncertainity in μ

- For the Bayesian setting for the multivariate case with fixed Σ

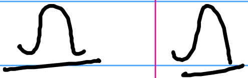
$$\mathbf{x} \sim \mathcal{N}(\mu, \Sigma),\ \ \mu \sim \mathcal{N}(\mu_0, \Sigma_0)\ \&\ p(\mathbf{x}|D) \sim \mathcal{N}(\mu_m, \Sigma_m)$$

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

- We now conclude our discussion on Bayesian Linear Regression..

$$w_i \sim N\left(0, \frac{1}{\lambda}\right)$$

$$[w_1 \cdots w_n] = \bar{w} = N\left(0, \frac{1}{\lambda} I\right) \hookrightarrow \Sigma = \begin{bmatrix} \frac{1}{\lambda} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{\lambda} \end{bmatrix}$$



$w_i \perp\!\!\!\perp w_j$ (w_i is independent of w_j)

$$Cov(w_i, w_j) = 0$$

Prior Distribution for w for Linear Regression

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

- We saw that when we try to maximize log-likelihood we end up with $\hat{\mathbf{w}}_{MLE} = (\Phi^T\Phi)^{-1}\Phi^T\mathbf{y}$
- We can use a Prior distribution on $\mathbf{w}$ to avoid over-fitting

  $$w_i \sim \mathcal{N}(0, \frac{1}{\lambda})$$

  ..Each component w_i is approximately bounded within $\pm\frac{3}{\sqrt{\lambda}}$.
  λ is also called the precision of the Gaussian
- Q1: How do deal with Bayesian Estimation for Gaussian distribution?
- Q2: Then what is the (collective) prior distribution of the n-dimensional vector $\mathbf{w}$?

Multivariate Normal Distribution and MAP estimate

Recall: $w = N(\mu, \Sigma) = \frac{1}{(2\pi|\Sigma|)^{n/2}} e^{-\frac{1}{2}(w-\mu)^T \Sigma^{-1}(w-\mu)}$

$\Sigma^{-1} = \lambda I, |\Sigma| = \frac{1}{\lambda^n}$

1. If $w_i \sim \mathcal{N}(0, \frac{1}{\lambda})$ then $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \frac{1}{\lambda} I)$ where I is an $n \times n$ identity matrix
2. $\Rightarrow$ That is, $\mathbf{w}$ has a multivariate Gaussian distribution $\Pr(\mathbf{w}) = \frac{1}{(\frac{2\pi}{\lambda})^{\frac{n}{2}}} e^{-\frac{\lambda}{2}\|w\|_2^2}$ with $\mu_0 = \mathbf{0}$. $\Sigma_0 = \frac{1}{\lambda} I$
3. We will specifically consider Bayesian Estimation for multivariate Gaussian (Normal) Distribution on $\mathbf{w}$: $\frac{1}{(2\pi)^{\frac{n}{2}}(\frac{1}{\lambda})^{\frac{1}{2}}} e^{-\frac{\lambda}{2}\|w\|_2^2}$

Substitute for $\Sigma_0 = \frac{1}{\lambda} I$

$[\epsilon_1 .. \epsilon_m] \sim N(0, \underbrace{\sigma^2 I}_{\to \Sigma})$

& determine $\Pr(w|D)$

$\Bigg\}$ H/w: