Introduction to Machine Learning - CS725

Instructor: Prof. Ganesh Ramakrishnan

Lecture 5 - 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:* $\hat{p} = \underset{p}{\operatorname{argmax}} \ {}^{n}C_h p^h (1-p)^{n-h} = \frac{h}{n}$
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 $\tilde{p} = \underset{H}{\operatorname{argmax}} \Pr(H \mid \mathcal{D}) = \underset{p}{\operatorname{argmax}} \Pr(p \mid \mathcal{D})$ $= \underset{p}{\operatorname{argmax}} \, Beta(p; \alpha + h, \beta + n - h) = \dfrac{\alpha + h - 1}{\alpha + \beta + n - 2}$

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
- 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!

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

$$w_i \sim \mathcal{N}(0, \tfrac{1}{\lambda}) \quad |w_i| \leq \frac{3}{\sqrt{\lambda}}$$

  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?

Q: How to estimate distr for w & $w|D$

$Pr(w)$ $Pr(w|D)$

$$Pr(w) = Pr(w_1, w_2 \ldots w_n)$$

$\sigma^2 = \frac{1}{\lambda}$

$$\text{s.t.} \quad Pr(w_i) = \frac{\sqrt{\lambda}}{\sqrt{2\pi}} e^{-\lambda w_i^2/2} = \mathcal{N}\left(0, \frac{1}{\lambda}\right)$$

$$\underbrace{Pr(w)}_{} = \mathcal{N}\left([0, \ldots 0], \frac{1}{\lambda}\begin{bmatrix} 1 & & 0 \\ & 1 & \\ 0 & & 1 \end{bmatrix}\right)$$

Density fn shown as concentric circles in 2-D

Q: What to do abt $Pr(w|D)$?

We will start with the simpler setting:

$$X \sim N(\mu, \sigma^2)$$

$$Pr(X|D) \sim ?$$

Evolve tricks in simpler setting

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

Interested in Posterior: $\underline{\Pr(\mu \mid \mathcal{D})}$

- We will temporarily generalize the discussion with x taking the place of ε and μ taking the place of w_i $\rightarrow w$
- y 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 $\underline{\Pr(\mu) = \mathcal{N}(\mu_0, \sigma_0^2)}$. Then the **posterior** is?

$$(\mu_{MLE}, \sigma^2_{MLE}) = \arg\max_{\mu, \sigma^2} \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^m \prod_{i=1}^{m} \exp\left(\frac{-1}{2\sigma^2}(x_i - \mu)^2\right)$$

How: Set $\frac{\partial}{\partial \mu} = 0$ & $\frac{\partial}{\partial \sigma^2} = 0$

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

$$P(\mu|\underbrace{x_1\ldots x_m}) = \frac{P(x_1\ldots x_m|\mu)\,P(\mu)}{\underbrace{P(x_1\ldots x_m)}}$$

Will be accounted for through normalization factor

$$= \frac{1}{Z} P(x_1\ldots x_m|\mu)\,P(\mu)$$

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

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

Detailed derivation

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

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

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

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

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

Trick: Determine params by matching coefficients of random variable (μ)

1 2

Match: Coefficients of μ, μ^2 (Random var)

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

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

ie

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

→ Precision of posterior increases as m 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:

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

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

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

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

$\sum x_i$ is $m\mu_{MLE}$

σ_m^2 already computed

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

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

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

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

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

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

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

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 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

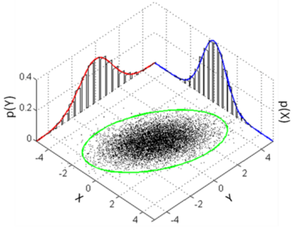
Summary: Conjugate Prior for (univariate) Gaussian

- Let $\Pr(X) \sim \mathcal{N}(\mu, \sigma^2)$ and let the data $\mathcal{D} = x_1 \ldots x_m$
- $\mu_{MLE} = \frac{1}{m}\sum_{i=1}^{m} x_i$ 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: $\Pr(\mu | x_1 \ldots x_m) = \mathcal{N}(\mu_m, \sigma^2_m)$ such that
- $\mu_m = (\frac{\sigma^2}{m\sigma_0^2 + \sigma^2}\mu_0) + (\frac{m\sigma_0^2}{m\sigma_0^2 + \sigma^2}\hat{\mu}_{ML})$ } → weighted mean of μ_0 & μ_{MLE} is μ_m.
- $\frac{1}{\sigma^2_m} = \frac{1}{\sigma_0^2} + \frac{m}{\sigma^2_{MLE}}$ } weighted harmonic mean of σ_0^2 & σ^2_{MLE} is σ^2_m

As $m \to \infty$, $\mu_m \to \mu_{MLE}$, $\sigma^2_m \to 0$

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



$\mu\in\Re^n$

$$\frac{e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}}{(2\pi)^{n/2}|\Sigma|^{1/2}}$$

$(\Sigma_{MLE})_{ij}$ $\propto$ correlation of ϕ_i with ϕ_j averaged across at data pts

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

Similar to univariate

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

$\sum[\ ]\,[\ ]=\sum[(\ )]$ = sum of rank 1 matrices

Summary for MAP estimation with Normal Distribution

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

Generalizing the inverse of a 1×1 matrix

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

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

} Univariate case

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

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

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

$$\Sigma_m^{-1}\mu_m = m\Sigma_{mle}^{-1}\mu_{mle} + \Sigma_0^{-1}\mu_0$$

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

For each j: $\sum_i (\mu_m)_i (\Sigma_m^{-1})_{ji} = m \sum_i (\mu_{mle})_i (\Sigma_{mle}^{-1})_{ji}$

Look at projection of mean on corre-lation of ϕ_j with all vars.

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..

Prior Distribution for w for Linear Regression

$Pr(w) = \mathcal{N}(w \mid 0, \frac{1}{\lambda} I)$ ✓

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

Goal!
$Pr(w \mid D)$?

- 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? ✓ Done!
- Q2: Then what is the (collective) prior distribution of the n-dimensional vector $\mathbf{w}$?

✓ Done

Multivariate Normal Distribution and MAP estimate

Recall. $x \sim \mathcal{N}(\mu, \Sigma)$ $\mu \sim \mathcal{N}(\mu_0, \Sigma_0)$ $Y = \Phi w + \epsilon$ [!]

Here: $\epsilon \sim \mathcal{N}(0, \sigma^2)$ $w \sim \mathcal{N}(0, \frac{1}{\lambda} I)$ $y = w^T \phi(x) + \epsilon$ (*)

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 $\mu_0 = \mathbf{0}$. $\Sigma_0 = \frac{1}{\lambda} I$ → $(w_i \perp\!\!\!\perp w_j)$
3. We will specifically consider Bayesian Estimation for multivariate Gaussian (Normal) Distribution on $\mathbf{w}$: $\frac{1}{(2\pi)^{\frac{n}{2}} (\frac{1}{\lambda})^{\frac{1}{2}}} e^{-\frac{\lambda}{2}\|w\|_2^2}$

$\Sigma_0 = \frac{1}{\lambda} I$ $\mu_0 = 0$ $\Pr(w|D) = \mathcal{N}(\mu_m, \Sigma_m)$

$\Sigma_m^{-1} = \lambda I + \underbrace{\Phi^T \Phi}_{} \times \frac{1}{\sigma^2}$ $\Sigma_m^{-1} \mu_m = \Sigma_0^{-1} \mu_0 + \Phi^T Y / \sigma^2$

Surprisingly independent of Y! (owing to linear eq (*)

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 $\hat{\mathbf{w}}_{MLE} = (\Phi^T\Phi)^{-1}\Phi^T\mathbf{y}$
- We can use a Prior distribution on w to avoid over-fitting
$$w_i \sim \mathcal{N}(0, \frac{1}{\lambda})$$
(that is, each component w_i is approximately bounded within $\pm\frac{1}{\sqrt{\lambda}}$ by the $3-\sigma$ rule)
- We want to find $P(\mathbf{w}|D) = \mathcal{N}(\mu_m, \Sigma_m)$
Invoking the Bayes Estimation results from before:

$$\frac{1}{|\Sigma_m|^{1}(2\pi)^{m/2}} e^{-w^T w \cdot \frac{1}{\lambda}} = P(w)$$

$$P(w|D) \propto P(w) \cdot \underbrace{P(D|w)}_{\prod_i \left(\frac{1}{\sigma\sqrt{2\pi}}\right) e^{-\frac{1}{2\sigma^2}(y - w^T\phi(x_i))^2}} \ldots$$

[Do coeff matching]

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 $\hat{\mathbf{w}}_{MLE} = (\Phi^T \Phi)^{-1} \Phi^T \mathbf{y}$
- We can use a Prior distribution on w to avoid over-fitting
$$w_i \sim \mathcal{N}(0, \tfrac{1}{\lambda})$$
(that is, each component w_i is approximately bounded within $\pm \frac{1}{\sqrt{\lambda}}$ by the $3 - \sigma$ rule)
- We want to find $P(\mathbf{w}|D) = \mathcal{N}(\mu_m, \Sigma_m)$
Invoking the Bayes Estimation results from before:
$$\Sigma_m^{-1} \mu_m = \Sigma_0^{-1} \mu_0 + \Phi^T y / \sigma^2$$
$$\Sigma_m^{-1} = \Sigma_0^{-1} + \frac{1}{\sigma^2} \Phi^T \Phi$$

Finding μ_m & Σ_m for $\mathbf{w}$

Setting $\Sigma_0 = \frac{1}{\lambda} I$ and $\mu_0 = \mathbf{0}$

$$
\begin{aligned}
\Sigma_m^{-1}\mu_m &= \Phi^T \mathbf{y}/\sigma^2 \\
\Sigma_m^{-1} &= \lambda I + \Phi^T \Phi/\sigma^2 \\
\mu_m &= \frac{(\lambda I + \Phi^T\Phi/\sigma^2)^{-1}\Phi^T \mathbf{y}}{\sigma^2}
\end{aligned}
$$

or

$$\mu_m = (\lambda\sigma^2 I + \Phi^T\Phi)^{-1}\Phi^T \mathbf{y}$$

MAP and Bayes Estimates

- $\Pr(\mathbf{w} \mid \mathcal{D}) = \mathcal{N}(\mathbf{w} \mid \mu_m, \Sigma_m)$
- The **MAP estimate** or mode under the Gaussian posterior is the mode of the posterior $\Rightarrow$

$$\hat{w}_{MAP} = \underset{\mathbf{w}}{\operatorname{argmax}} \mathcal{N}(\mathbf{w} \mid \mu_m, \Sigma_m) = \underline{\mu_m}$$

- Similarly, the **Bayes Estimate**, or the expected value under the Gaussian posterior is the mean $\Rightarrow$

$$\hat{w}_{Bayes} = E_{\Pr(\mathbf{w}|\mathcal{D})}[\mathbf{w}] = E_{\mathcal{N}(\mu_m, \Sigma_m)}[\mathbf{w}] = \underline{\mu_m}$$

- Summarily:

$$\mu_{MAP} = \mu_{Bayes} = \mu_m = (\lambda\sigma^2 I + \Phi^T\Phi)^{-1}\Phi^T\mathbf{y}$$

$$\Sigma_m^{-1} = \lambda I + \frac{\Phi^T\Phi}{\sigma^2}$$

From Bayesian Estimates to (Pure) Bayesian Prediction

	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	$P(y\|x, w, \frac{1}{\lambda} \cdots) = \mathcal{N}(\mu_0^T \phi(x), \cdots)$	$p(\theta\|D) = \frac{p(D\|\theta)p(\theta)}{\int_m p(D\|\theta)p(\theta)d\theta}$ $p(D\|\theta) = \prod_{i=1} p(x_i\|\theta)$ $p(x\|D) = \int_\theta p(x\|\theta)p(\theta\|D)d\theta$

where θ is the parameter