Introduction to Machine Learning - CS725
Instructor: Prof. Ganesh Ramakrishnan
Lecture 07 - Bayesian Linear Regression (Gaussian and Laplacian priors), Regularized Linear Regression (Ridge Regression and Lasso)

Recap: Summary for MAP estimation with Normal Distribution

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

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

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

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

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

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

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

Different Estimators

Recap: Mean and Mode coincide for (Multivariate) Gaussian ==> Bayes Estimate = MAP estimate

	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	$p(\theta\|D) \sim N(\mu_m, \Sigma_m)$ $p(x\|D) \sim N(\mu_m + \cdots, \Sigma_m + \cdots)$	$p(\theta\|D) = \frac{p(D\|\theta)p(\theta)}{\int p(D\|\theta)p(\theta)d\theta}$ $p(D\|\theta) = \prod_{i=1}^{m} p(x_i\|\theta)$ $p(x\|D) = \int_\theta p(x\|\theta)p(\theta\|D)$

Recap: 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!

Recap: 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

$\mu = 0$, since ϕ_i should have $w_i \approx 0$ unless ϕ_i provides "signal"

$$w_i \sim \mathcal{N}(0, \frac{1}{\lambda}) \rightarrow \frac{1}{\lambda} = \sigma^2 = \frac{1}{\text{precision}}$$

(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?

Recap: Multivariate Normal Distribution and MAP estimate

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. Consider Bayesian Estimation for multivariate Gaussian on $\mathbf{w}$

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

$\mu_0 = 0$

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

- Given $y = \mathbf{w}^T \phi(\mathbf{x}) + \varepsilon$ and $\varepsilon \sim \mathcal{N}(0, \sigma^2) \Rightarrow$ $y \sim \mathcal{N}(\mathbf{w}^T\phi(\mathbf{x}), \sigma^2)$, $\mathbf{w} \sim \mathcal{N}(\mu_0, \Sigma_0)$ where, $w_i \sim \mathcal{N}(0, \frac{1}{\lambda})$
- We want to find $P(\mathbf{w}|D) = \mathcal{N}(\mu_m, \Sigma_m)$
  Invoking the Bayes Estimation results from before (homework):

$$\Pr(w|D) \propto P(D|w)\,P(w) \propto \prod_{i=1}^{m} \exp\left(-\left(\frac{y_i - w^T\phi(x_i)}{2\sigma^2}\right)^2\right) * \exp\left[-(w-\mu_0)^T\Sigma_0^{-1}(w-\mu_0)\right]$$

$$= \exp\left[-w^T\left(\Sigma_0^{-1} + \sum_{i=1}^{m}\underbrace{\frac{\phi^T(x_i)\phi(x_i)}{2\sigma^2}}\right)w^T + 2\sum_{i=1}^{m} w^T\left(\underbrace{\frac{y_i\phi(x_i)}{2\sigma^2}} + \mu_0^T\Sigma_0^{-1}\right) - \ldots\right]$$

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

- Given $y = \mathbf{w}^T\phi(\mathbf{x}) + \varepsilon$ and $\varepsilon \sim \mathcal{N}(0, \sigma^2) \Rightarrow$
  $y \sim \mathcal{N}(\mathbf{w}^T\phi(\mathbf{x}), \sigma^2)$, $\mathbf{w} \sim \mathcal{N}(\mu_0, \Sigma_0)$ where, $w_i \sim \mathcal{N}(0, \frac{1}{\lambda})$
- We want to find $P(\mathbf{w}|D) = \mathcal{N}(\mu_m, \Sigma_m)$
  Invoking the Bayes Estimation results from before (homework):

$\mu_0 = 0$

$\underbrace{\Sigma_0^{-1} = \lambda I}$ → Substitute !

$$\Sigma_m^{-1} = \frac{1}{\sigma^2}\Phi^T\Phi + \Sigma_0^{-1}$$
$$\Sigma_m^{-1}\mu_m = \Phi^T y/\sigma^2 + \cancel{\Sigma_0^{-1}}\mu_0$$

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

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

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

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

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 $\Rightarrow$

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

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

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

- Summarily:

Recall: w_{MLE} had no σ^2 (& of course no λ)

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

$\Phi^T\Phi$

Predictive distribution for linear Regression

Pure Bayesian: $Pr(y|x, D)$

- $\hat{\mathbf{w}}_{MAP}$ helps avoid overfitting as it takes regularization into account
- But we miss the modeling of uncertainty when we consider only $\hat{\mathbf{w}}_{MAP}$
- **Eg:** While predicting diagnostic results on a new patient x, along with the value y, we would also like to know the uncertainty of the prediction $\Pr(y \mid x, D)$. Recall that $y = \mathbf{w}^T\phi(x) + \varepsilon$ and $\varepsilon \sim \mathcal{N}(0, \sigma^2)$

$$\Pr(y \mid \mathbf{x}, \mathcal{D}) = \Pr(y \mid \mathbf{x}, < \mathbf{x}_1, y_1 > \ldots < \mathbf{x}_m, y_m >)$$

Expect: $Pr(y|x, D) \sim N(\mu_m^T\phi(x), \phi^T(x)\Sigma_m\phi(x) + \sigma^2)$

Pure Bayesian Regression Summarized (optional)

By definition, regression is about finding $(y \mid \mathbf{x}, \mathcal{D})$. By Bayes Rule

$$\begin{aligned}
\Pr(y \mid \mathbf{x}, \mathcal{D}) &= \Pr(y \mid \mathbf{x}, < \mathbf{x}_1, y_1 > \ldots < \mathbf{x}_m, y_m >) \\
&= \int_{\mathbf{w}} \Pr(y|\mathbf{w}; \mathbf{x}) \Pr(\mathbf{w} \mid \mathcal{D}) d\mathbf{w} \\
&\sim \mathcal{N}\left(\mu_m^T \phi(\mathbf{x}), \sigma^2 + \phi^T(\mathbf{x})\Sigma_m \phi(\mathbf{x})\right)
\end{aligned}$$

where

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

$$\mathbf{w} \sim \mathcal{N}(0, \alpha I) \text{ and } \mathbf{w} \mid \mathcal{D} \sim \mathcal{N}(\mu_m, \Sigma_m)$$

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

$$\text{Finally} \quad y \sim \mathcal{N}(\mu_m^T \phi(\mathbf{x}), \phi^T(\mathbf{x})\Sigma_m \phi(\mathbf{x}))$$

MAP (and Bayes) Inference (Rewriting differently)

$$\mathbf{w}_{MAP} = \underset{\mathbf{w}}{\operatorname{argmax}} \ \Pr(\mathbf{w} \mid \mathcal{D}) = \underset{\mathbf{w}}{\operatorname{argmax}} \ \log \Pr(\mathbf{w} \mid \mathcal{D}), \text{ where,}$$

$$-\log \Pr(\mathbf{w} \mid \mathcal{D}) = \frac{n}{2}\log(2\pi) + \frac{1}{2}\log|\Sigma_m| + \frac{1}{2}(\mathbf{w}-\mu_m)^T \Sigma_m^{-1}(\mathbf{w}-\mu_m)$$

$$\mathbf{w}_{MAP} = \underset{\mathbf{w}}{\operatorname{argmin}} - \log \Pr(\mathbf{w}) = \underset{\mathbf{w}}{\operatorname{argmin}} \frac{1}{2}\mathbf{w}^T \Sigma_m^{-1}\mathbf{w} - \mathbf{w}^T \Sigma_m^{-1}\mu_m$$

(expanding/canceling redundant terms & completing squares: Tut 3) → By substituting for Σ_m & μ_m

Recall: log is monotonically increasing

$$\underset{w}{\operatorname{argmax}} f(w) = \underset{w}{\operatorname{argmin}} -f(w)$$

MAP (and Bayes) Inference

$$\mathbf{w}_{MAP} = \underset{\mathbf{w}}{\operatorname{argmax}} \; \Pr(\mathbf{w} \mid \mathcal{D}) = \underset{\mathbf{w}}{\operatorname{argmax}} \; \log \Pr(\mathbf{w} \mid \mathcal{D}), \text{ where,}$$

$$-\log \Pr(\mathbf{w} \mid \mathcal{D}) = \frac{n}{2}\log(2\pi) + \frac{1}{2}\log|\Sigma_m| + \frac{1}{2}(\mathbf{w}-\mu_m)^T\Sigma_m^{-1}(\mathbf{w}-\mu_m)$$

$$\mathbf{w}_{MAP} = \underset{\mathbf{w}}{\operatorname{argmin}} - \log \Pr(\mathbf{w}) = \underset{\mathbf{w}}{\operatorname{argmin}} \frac{1}{2}\mathbf{w}^T\Sigma_m^{-1}\mathbf{w} - \mathbf{w}^T\Sigma_m^{-1}\mu_m$$

(expanding/canceling redundant terms & completing squares: Tut 3)

$$\boxed{\mathbf{w}_{MAP} = \underset{\mathbf{w}}{\operatorname{argmin}} \frac{1}{2}\underbrace{||\phi\mathbf{w} - \mathbf{y}||^2}_{\text{Least squares loss}} + \underbrace{\sigma^2\lambda||\mathbf{w}||^2}_{\text{Regularizer}} = \mathbf{w}_{Ridge}}$$

is the same as that of Penalized *Regularized Regression*.

$$w_{MAP} = \underset{w}{\operatorname{argmin}} \ \frac{1}{2}\|\Phi w - y\|^2 + \lambda\sigma^2\|w\|^2$$

Independently $\|w\|^2$ is minimized only when each $w_i = 0$

THIS IS COMMON SENSE! ALL WE HAVE DONE IS GIVEN SOME PROBABILISTIC INTERPRETATION TO COMMON SENSE!

Penalized Regularized Least Squares Regression

- The Bayes and MAP estimates for Linear Regression coincide with *Regularized Ridge Regression*

$$\mathbf{w}_{Ridge} = \underset{\mathbf{w}}{\arg\min}\ ||\Phi\mathbf{w} - \mathbf{y}||_2^2 + \lambda||\mathbf{w}||_2^2$$

Eg: If $\phi_i = \phi_j$, $w_i \neq 0$ & want $w_j = 0$

- **Intuition:** To discourage redundancy and/or stop coefficients of $\mathbf{w}$ from becoming too large in magnitude, add a penalty to the error term used to estimate parameters of the model.
- The general **Penalized Regularized L.S Problem**:

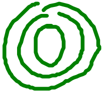
$$\mathbf{w}_{Reg} = \underset{\mathbf{w}}{\arg\min}\ ||\Phi\mathbf{w} - \mathbf{y}||_2^2 + \lambda\Omega(\mathbf{w})$$

Can also be generic Error(ϕ, w, y)

Penalized Regularized Least Squares: Examples

- The general **Penalized Regularized L.S Problem**:

$$\mathbf{w}_{Reg} = \arg\min_{\mathbf{w}} \; ||\Phi\mathbf{w} - \mathbf{y}||_2^2 + \lambda\Omega(\mathbf{w})$$



corners tend to attract more solns

- $\Omega(\mathbf{w}) = ||\mathbf{w}||_2^2 \Rightarrow$ **Ridge Regression**
- $\Omega(\mathbf{w}) = ||\mathbf{w}||_1 \Rightarrow$ **Lasso** (A compromise)
- $\Omega(\mathbf{w}) = ||\mathbf{w}||_0 \Rightarrow$ **Support-based penalty** → # non-zeros..

not its / differentiable

- Some $\Omega(\mathbf{w})$ correspond to priors that can be expressed in close form. Some give good working solutions. Some norms are mathematically easier to handle

$||w||_2^2 \sim N(\cdot) \sim \exp(||\cdot||^2)$. $||w||_1 \sim Lap(\cdot) \sim \exp(|\cdot|)$

Constrained Regularized Least Squares Regression

- **Intuition:** To discourage redundancy and/or stop coefficients of $\mathbf{w}$ from becoming too large in magnitude, constrain the error minimizing estimate using a penalty
- The general **Constrained Regularized L.S. Problem**:

$$\mathbf{w}_{Reg} = \arg\min_{\mathbf{w}} \; ||\Phi\mathbf{w} - \mathbf{y}||_2^2$$

$$\textit{such that } \Omega(\mathbf{w}) \leq \theta$$

- Claim: For any Penalized formulation with a particular λ, there exists a corresponding Constrained formulation with a corresponding θ
- **Proof of Equivalence:** Requires tools of Optimization/duality

Constrained Regularized Least Squares: Examples

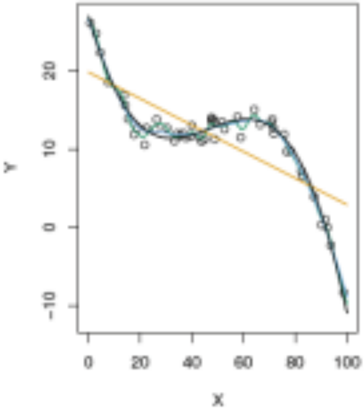
- The general **Constrained Regularized L.S. Problem**:

$$\mathbf{w}_{Reg} = \arg\min_{\mathbf{w}} \; ||\Phi\mathbf{w} - \mathbf{y}||_2^2$$

$$\textit{such that } \Omega(\mathbf{w}) \leq \theta$$

  - $\Omega(\mathbf{w}) = ||\mathbf{w}||_2^2 \Rightarrow$ **Ridge Regression**
  - $\Omega(\mathbf{w}) = ||\mathbf{w}||_1 \Rightarrow$ **Lasso**
  - $\Omega(\mathbf{w}) = ||\mathbf{w}||_0 \Rightarrow$ **Support-based penalty**

Polynomial regression



- Consider a degree 3 polynomial regression model as shown in the figure
- Each bend in the curve corresponds to increase in $\|w\|$
- Eigen values of $(\Phi^\top\Phi + \lambda I)$ are indicative of curvature. Increasing λ reduces the curvature

Do Closed-form solutions Always Exist?

- Linear regression and Ridge regression both have closed-form solutions
  - For linear regression,

$$w^* = (\Phi^\top \Phi)^{-1} \Phi^\top y$$

  - For ridge regression,

$$w^* = (\Phi^\top \Phi + \lambda I)^{-1} \Phi^\top y$$

  (for linear regression, $\lambda = 0$)

- What about optimizing the formulations (constrained/penalized) of Lasso (L_1 norm)? And support-based penalty (L_0 norm)?: Also requires tools of Optimization/duality

Lasso Regularized Least Squares Regression

- The general **Penalized Regularized L.S Problem**:

$$\mathbf{w}_{Reg} = \arg\min_{\mathbf{w}} \ ||\Phi\mathbf{w} - \mathbf{y}||_2^2 + \lambda\Omega(\mathbf{w})$$

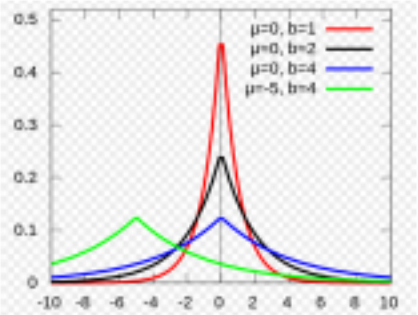
  - $\Omega(\mathbf{w}) = ||\mathbf{w}||_1 \Rightarrow$ **Lasso**
- *Lasso* Regression

$$\mathbf{w}_{lasso} = \arg\min_{\mathbf{w}} \ ||\Phi\mathbf{w} - \mathbf{y}||_2^2 + \lambda||\mathbf{w}||_1$$

- Lasso is the MAP estimate of Linear Regression subject to Laplace Prior on $\mathbf{w} \sim Laplace(0, \theta)$

$$Laplace(w_i \mid \mu, b) = \frac{1}{2b}\exp\left(-\frac{|w_i - \mu|}{b}\right)$$

Gaussian Hare vs. Laplacian Tortoise



- Gaussian easier to estimate

$$w^{Gaussian}_{MAP} = (\Phi^T\Phi + \lambda I)^{-1}\Phi^T y$$

- Laplacian yields more sparsity

No closed form for $w^{Lap.}_{MAP}$

Iterative algos...