Introduction to Machine Learning - CS725
Instructor: Prof. Ganesh Ramakrishnan
Lecture 7 - 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

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

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

- 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
$$w_i \sim \mathcal{N}(0, \frac{1}{\lambda})$$
(that is, each component w_i is approximately bounded within $\pm \frac{3}{\sqrt{\lambda}}$ 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:

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

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

- 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

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

  (that is, each component w_i is approximately bounded within $\pm \frac{3}{\sqrt{\lambda}}$ 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:

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

Finding μ_m & Σ_m for $\mathbf{w}$

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

$$\Sigma_m^{-1} \mu_m = \Phi^T \mathbf{y} / \sigma^2$$

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

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

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 $\Rightarrow$

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

- 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:

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

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

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		$p(\theta\|D) = \frac{p(D\|\theta)p(\theta)}{\int_m p(D\|\theta)p(\theta)d}$ $p(D\|\theta) = \prod_{i=1} p(x_i\|\theta)$ $p(x\|D) = \int_\theta p(x\|\theta)p(\theta\|D$

where θ is the parameter

Predictive distribution for linear Regression

- $\hat{\mathbf{w}}_{MAP}$ 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 $\varepsilon \sim \mathcal{N}(0, \sigma^2)$

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

Pure Bayesian Regression Summarized

- By definition, regression is about finding $(y \mid \mathbf{x}, < \mathbf{x}_1, y_1 > ... < \mathbf{x}_m, y_m >)$
- By Bayes Rule

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

where

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

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

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

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

Penalized Regularized Least Squares Regression

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

$$\mathbf{w}_{Ridge} = \arg\min_{\mathbf{w}} \; ||\Phi\mathbf{w} - \mathbf{y}||_2^2 + \lambda\sigma^2||\mathbf{w}||_2^2$$

- **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**:

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

  - $\Omega(\mathbf{w}) = ||\mathbf{w}||_2^2 \Rightarrow$ **Ridge Regression**
  - $\Omega(\mathbf{w}) = ||\mathbf{w}||_1 \Rightarrow$ **Lasso**
  - $\Omega(\mathbf{w}) = ||\mathbf{w}||_0 \Rightarrow$ **Support-based penalty**
- Some $\Omega(\mathbf{w})$ correspond to priors that can be expressed in close form. Some give good working solutions. However, for mathematical convenience, some norms are easier to handle

Constrained Regularized Least Squares Regression

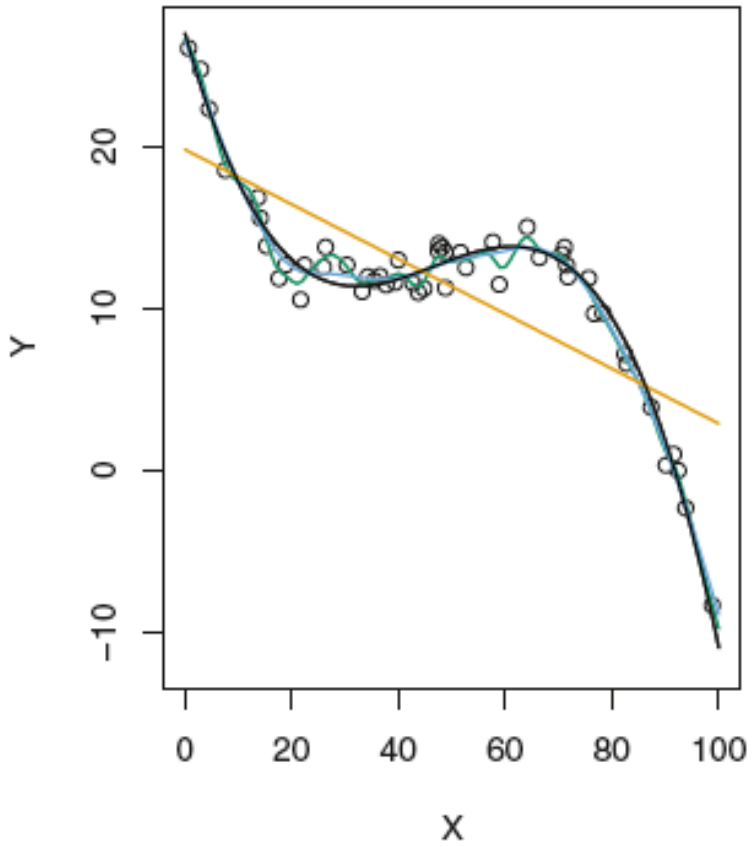
- **Intuition:** To discourage redundancy and/or stop coefficients of $\mathbf{w}$ from becoming too large in magnitude, constrain the error minimizing estimate using a penalty
- The general **Constrained Regularized L.S. Problem**:

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

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

- Claim: For any Penalized formulation with a particular λ, there exists a corresponding Constrained formulation with a corresponding θ
  - $\Omega(\mathbf{w}) = ||\mathbf{w}||_2^2 \Rightarrow$ **Ridge Regression**
  - $\Omega(\mathbf{w}) = ||\mathbf{w}||_1 \Rightarrow$ **Lasso**
  - $\Omega(\mathbf{w}) = ||\mathbf{w}||_0 \Rightarrow$ **Support-based penalty**
- **Proof of Equivalence:** Requires tools of Optimization/duality

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

Why is Lasso Interesting?

Support Vector Regression

One more formulation before we look at Tools of Optimization/duality