# 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

$$y = \mathbf{w}^T\phi(x) + \varepsilon$$

$$\varepsilon \sim \mathcal{N}(0, \sigma^2)$$

$$y \sim \mathcal{N}(w^T\phi(x), \sigma^2)$$

- 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

$$\mu_0 = 0 \leftarrow \underbrace{w_i \sim \mathcal{N}(0, \tfrac{1}{\lambda})}\rightarrow 1/\lambda = (\Sigma_0)_{ii}$$

  (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:

$$\underline{\Sigma^{-1} = \frac{1}{\sigma^2}\phi^T\phi} \ \& \ \Sigma_0 = \frac{1}{\lambda}I \Leftrightarrow \Sigma_m^{-1} = \Sigma_0^{-1} + m\Sigma^{-1}$$

$$\underline{\mu = 0} \ \& \ \mu_0 = 0 \qquad \Leftrightarrow \Sigma_m^{-1}\mu_m = m\Sigma^{-1}\hat{\mu}_{MLE} + \Sigma_0^{-1}\mu_0$$

$$w \sim \mathcal{N}(\mu_0, \Sigma_0) = \mathcal{N}(0, \tfrac{1}{\lambda} I)$$

$$P(w|D) \propto P(D|w)\, P(w)$$

$$\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^m \prod_{i=1}^{m} \exp\left[\frac{-1}{2\sigma^2}\left(y_i - w^T\phi(x_i)\right)^2\right] = \mathcal{N}(\mu, \Sigma)$$

$$= \qquad \left( ( \quad - \mu)^T \Sigma^{-1} ( \quad - \mu) \right)$$

Homework: Complete the derivation

# 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:

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

$$\left[ \Sigma^{-1} = \frac{1}{\sigma^2} \Phi^T \Phi \right]$$

# Finding $\mu_m$ & $\Sigma_m$ 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

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

If $\lambda = 0 \Rightarrow$ No prior belief abt $\lambda$

$$\mu_m = (0 + \phi^T \phi)^{-1}\phi^T y = (\phi^T \phi)^{-1}\phi^T y = w_{LS}^*$$

# 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 $\Rightarrow$

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

$\lambda \to 0 \Rightarrow \omega_{MAP} = \omega_{MLE}$

- 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:

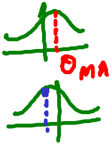
$$\mu_{MAP} = \mu_{Bayes} = \mu_m = (\lambda\sigma^2 I + \Phi^T\Phi)^{-1}\Phi^T\mathbf{y}$$

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

$$\phi = \begin{bmatrix} \phi(x_1) \\ \phi(x_2) \\ \phi(x_m) \end{bmatrix}$$

# From Bayesian Estimates to (Pure) Bayesian Prediction

Most important ↓

| | **Point?** | $p(x\|D)$ |
|---|---|---|
| MLE | $\hat{\theta}_{MLE} = \text{argmax}_\theta\, LL(D\|\theta)$ | $p(x\|\theta_{MLE}) = \mathcal{N}(\omega_{MLE}^T \phi(x), \sigma^{-2})$ |
| Bayes Estimator | $\hat{\theta}_B = E_{p(\theta\|D)} E[\theta]$ | $p(x\|\theta_B) = \mathcal{N}(\omega_B^T \phi(x), \sigma^{-2})$ |
| MAP | $\hat{\theta}_{MAP} = \text{argmax}_\theta\, p(\theta\|D)$ | $p(x\|\theta_{MAP}) = \mathcal{N}(\omega_{MAP}^T \phi(x), \sigma^{-2})$ |
| Pure Bayesian (Quite ambitious) |  $\theta_{MAP}$ $\theta_{MLE}$ posterior → likelihood → | $p(\theta\|D) = \frac{p(D\|\theta)p(\theta)}{\int_m p(D\|\theta)p(\theta)d\theta}$ <br> $p(D\|\theta) = \prod_{i=1}^{m} p(x_i\|\theta)$ <br> $p(x\|D) = \int_\theta p(x\|\theta)p(\theta\|D)\,d\theta$ |

where $\theta$ is the parameter

Averaging over all models/parameters

For Bernoulli with Beta prior:

$p_{MAP} = \frac{n+\alpha-1}{n+\alpha+\beta-2}$ $\quad p_{Bayes} = \frac{n+\alpha}{n+\alpha+\beta}$

# Predictive distribution for linear Regression

- $\hat{\mathbf{w}}_{MAP}$ helps avoid overfitting as it takes regularization into account  is a hack
- But we miss the modeling of uncertainty when we consider only $\hat{\mathbf{w}}_{MAP}$ : $P(y \mid x, \hat{w}_{MAP}) = \mathcal{N}(w_{MAP}^T \phi(x), \sigma^2)$ ?
- **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 $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ → factoring in $P(w|D)$

$x \to$ new pt
$\langle x_1, y_1 \rangle \ldots \langle x_m, y_m \rangle$ are old pts

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

$$= \int_w \underbrace{\Pr(y|x,w)}_{\to \mathcal{N}(w^T\phi(x), \sigma^2)} \underbrace{P(w|\langle x_1 y_1\rangle \ldots \langle x_m y_m\rangle)}_{\to \mathcal{N}(\mu_m, \Sigma_m)} dw$$

# Pure Bayesian Regression Summarized

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

x = Not a r.v
& hence preceded by ";"

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

*where*

$$y = \mathbf{w}^T \phi(\mathbf{x}) + \varepsilon \text{ and } \varepsilon \sim \mathcal{N}(0, \sigma^2)$$
$$\mathbf{w} \sim \mathcal{N}(0, \alpha I) \text{ and } \mathbf{w} \mid \mathcal{D} \sim \mathcal{N}(\mu_m, \Sigma_m)$$
$$\mu_m = (\lambda\sigma^2 I + \Phi^T\Phi)^{-1}\Phi^T \mathbf{y} \text{ and } \Sigma_m^{-1} = \lambda I + \Phi^T\Phi/\sigma^2$$
$$\text{Finally } \; y \sim \mathcal{N}(\mu_m^T \phi(\mathbf{x}), \phi^T(\mathbf{x}) \Sigma_m \phi(\mathbf{x}))$$

$+ \sigma^2$

# Penalized Regularized Least Squares Regression

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

[Ridge regression]

By taking gradient & setting it to 0

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

- **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**:

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

Abt optimization

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

$\left(\sum |w_i|^p\right)^{1/p} = \|w\|_p$

$p \to 0 \Rightarrow \|w\|_p =$ # non zero components

- 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**:

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

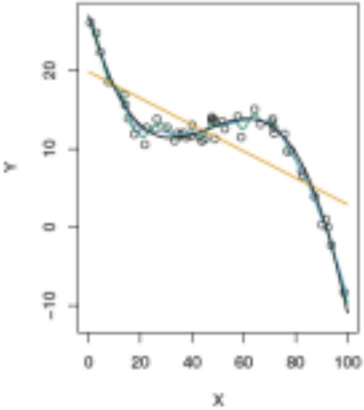
$$\textit{such that } \Omega(\mathbf{w}) \leq \theta$$

- Claim: For any Penalized formulation with a particular $\lambda$, there exists a corresponding Constrained formulation with a corresponding $\theta$
  - $\Omega(\mathbf{w}) = ||\mathbf{w}||_2^2 \Rightarrow$ **Ridge Regression**
  - $\Omega(\mathbf{w}) = ||\mathbf{w}||_1 \Rightarrow$ **Lasso**
  - $\Omega(\mathbf{w}) = ||\mathbf{w}||_0 \Rightarrow$ **Support-based penalty**

: we will later see $\theta \leftrightarrow \lambda$ mapping for ridge regression

- **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 $\lambda$ reduces the curvature

$\phi(x) = [x, x^2 \, .. ]$

$w_{ridge} = (\phi^T\phi + \lambda I)^{-1}\phi^T y$

Let $\lambda_1 \ldots \lambda_n$ be eigenvalues of $\phi^T\phi$

Then $(\phi^T\phi + \lambda I)$ has eigenvalues

$\lambda_1+\lambda, \lambda_2+\lambda, \ldots \lambda_n+\lambda$

If $\min(\lambda_i)$ was close to $0$, matrix $\phi^T\phi$ can be tricky to invert (low condition #)

$\Rightarrow (\phi^T\phi + \lambda I) \rightarrow$ more robustly non-singular (improved condition number)

# 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,

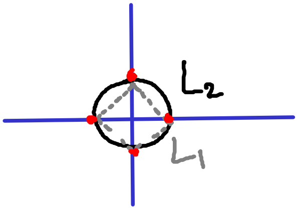
$$w^* = (\Phi^\top \Phi + \lambda I)^{-1} \Phi^\top y$$

  (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?

$L_0 \Rightarrow$ # non-zero components




Level curves of $L_1$, $L_2$

Level curves for $L_1$ tend to have more corners (•) than $L_2$ and hence, tend to yield sparser solutions

ie more $w_i$'s $= 0$

# Support Vector Regression

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