Introduction to Machine Learning - CS725

Instructor: Prof. Ganesh Ramakrishnan

Lecture 6 - Support Vector Regression and Optimization Basics

From Bayesian Estimates to (Pure) Bayesian Prediction

Special case by sampling

	Point?	$p(x \mid D)$
MLE	$\hat{\theta}_{MLE} = \text{argmax}_\theta\, LL(D \mid \theta)$	$p(x \mid \theta_{MLE})$
Bayes Estimator	$\hat{\theta}_B = E_{p(\theta \mid D)} E[\theta]$	$p(x \mid \theta_B)$
MAP	$\hat{\theta}_{MAP} = \text{argmax}_\theta\, p(\theta \mid D)$	$p(x \mid \theta_{MAP})$
Pure Bayesian	often approximated by sampling highly probable $p(\theta \mid D)$	$p(\theta \mid D) = \dfrac{p(D \mid \theta)p(\theta)}{\int_m p(D \mid \theta)p(\theta)d\theta}$ $p(D \mid \theta) = \prod_{i=1} p(x_i \mid \theta)$ $p(x \mid D) = \int_\theta p(x \mid \theta)p(\theta \mid D)d\theta$

} Density on query pt x using a param estimate

where θ is the parameter

$$p(x \mid D) = \int_\theta p(x \mid \theta)\, p(\theta \mid D)\, d\theta$$

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

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

$$= \int_w \underbrace{P(y \mid x, w)} \; \underbrace{P(w \mid D)}_{\mathcal{N}(\mu_m, \Sigma_m)} \, dw$$

Pure Bayesian Regression Summarized

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

MGF is $\exp(\mu t + \frac{\sigma^2 t^2}{2})$ & $\Phi(X+Y) = \Phi(X)\Phi(Y)$

$X \sim N(\mu_1, \sigma_1^2)$

$\perp$

$Y \sim N(\mu_2, \sigma_2^2)$

$X+Y \sim N(\mu_1+\mu_2, \sigma_1^2+\sigma_2^2)$

We have already seen a result for product!

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

$$= \int_{\mathbf{w}} \Pr(y|\mathbf{w}; \mathbf{x}) \Pr(\mathbf{w} \mid \mathcal{D}) d\mathbf{w}$$

$$\sim \mathcal{N}\left(\mu_m^T \phi(\mathbf{x}), \sigma^2 + \phi^T(\mathbf{x}) \Sigma_m \phi(\mathbf{x})\right)$$

$v^T A v$, $A \geq 0$ $\Rightarrow$ will be non-neg

$\langle \phi(x), E(w|D) \rangle$

where

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

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

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

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

MAP (and Bayes) Inference

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

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

$$\Pr(w \mid D) = \frac{1}{(2\pi)^{n/2}} \frac{1}{|\Sigma|^{1/2}} \exp\left(\frac{1}{2}(w - \mu_m)^T \Sigma_m^{-1} (w - \mu_m)\right)$$

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

..... (expanding & canceling out redundant terms & completing squares: Tutorial 3)

$\equiv \max_{w}$ - Least squares objective - ...

$\circledast$ can be ignored

$$w^T \Sigma_m^{-1} w = \mathrm{tr}(\Sigma_m^{-1}) w^T w = f(\Phi^T \Phi, \ldots)$$

Story so far: Least sq reg

Probabilistic interpretation

MLE & Bayesian estimation → $w_i \sim \mathcal{N}(0, \frac{1}{\lambda})$ view ①

$\parallel$ MAP $\parallel$ Bayes

Least sq.

Least sq + - - - ??

$$\min_w \|\Phi w - y\|_2^2$$

$$\min_w \|\Phi w - y\|_2^2 + \underbrace{\lambda \|w\|_2^2}$$ → view ②

New Discovery: Gaussian Prior ⇒ Penalizing $\|w\|_2$

MAP (and Bayes) Inference

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

$\epsilon \sim N(0, \sigma^2)$

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

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

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

..... (expanding & canceling out redundant terms & completing squares: Tutorial 3)

$$\mathbf{w}_{MAP} = \underset{\mathbf{w}}{\text{argmax}} \frac{1}{2\sigma^2}\mathbf{w}^T\left(\phi^T\phi\mathbf{w} - 2\phi^T\mathbf{y}\right) + \lambda\mathbf{w}^T\mathbf{w} = \underset{\mathbf{w}}{\text{argmax}} \frac{1}{2}||\phi\mathbf{w} - \mathbf{y}||^2 + \sigma^2\lambda||\mathbf{w}||^2 = \mathbf{w}_{Ridge}$$

is the same as that of *Regularized Regression*.

$$\mathbf{w}_{Ridge} = \underset{\mathbf{w}}{\text{argmin}}\ ||\phi\mathbf{w} - \mathbf{y}||_2^2 + \lambda\sigma^2||\mathbf{w}||_2^2$$

Penalized Regularized Least Squares Regression

Since we want many w_i's $=0$, we don't want $p<0$

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

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

Minimize error but while controlling $\|w\|$

- **Intuition:** To discourage redundancy and/or stop coefficients of **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**:

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

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

$p=2 \rightarrow$
- $\Omega(\mathbf{w}) = ||\mathbf{w}||_2^2 \Rightarrow$ **Ridge Regression** $= \left(\sum_i w_i^2\right)$

$p=1 \rightarrow$
- $\Omega(\mathbf{w}) = ||\mathbf{w}||_1 \Rightarrow$ **Lasso** $= \sum |w_i|$

$p \rightarrow 0$
- $\Omega(\mathbf{w}) = ||\mathbf{w}||_0 \Rightarrow$ **Support-based penalty** $\rightarrow$ combinatorial

$p \rightarrow 0 \Rightarrow$ nonzero # of w_i's

- 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

Questions to answer:

① Can $\|w\|_1$ approximately achieve the objective of $\|w\|_0$?

To actually solve

$$w_{\text{lasso}} = \arg\min \|\phi w - y\|^2 + \lambda \|w\|_1$$

we need optimization algos.

$p \in [0, 1, 2, \ldots \infty)$

desirable

Not too effective

Computationally burdensome to have to compute x^5 even though $w_5 \to 0$

② Is there a probabilistic interpretation to $\|w\|_1$

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**:

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

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

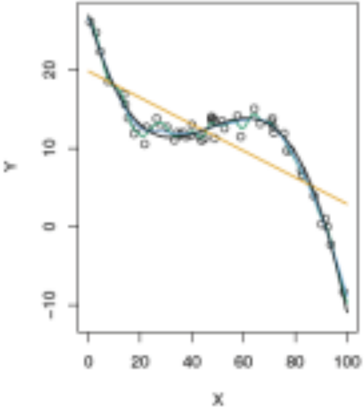
- Claim: For any Penalized formulation with a particular λ, there exists a corresponding Constrained formulation with a corresponding θ
  - $\Omega(\mathbf{w}) = ||\mathbf{w}||_2^2 \Rightarrow$ **Ridge Regression**
  - $\Omega(\mathbf{w}) = ||\mathbf{w}||_1 \Rightarrow$ **Lasso**
  - $\Omega(\mathbf{w})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

$\}\ \exists\ \theta = f(\lambda)$

- **Proof of Equivalence:** Requires tools of Optimization/duality

An implicit goal in regularization: Give user freedom to overestimate # of features

Polynomial regression



Objective: $\|\Phi w - y\|^2 + \lambda \|w\|_2^2$

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

eigenvalues → curvature

- 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

Tutorial 3

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,

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

Analysis Sufficient

(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

Algorithms required!

Lasso Regularized Least Squares Regression

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

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

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

- *Lasso* Regression

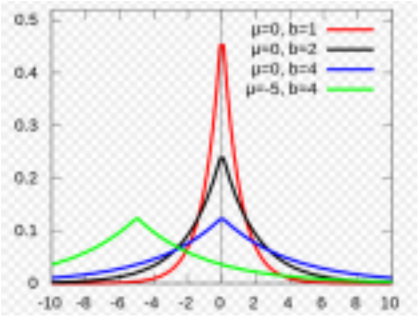
$$\mathbf{w}_{lasso} = \arg\min_{\mathbf{w}} \ ||\Phi\mathbf{w} - \mathbf{y}||_2^2 + \lambda||\mathbf{w}||_1^2$$

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

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

determines curvature

Gaussian Hare vs. Laplacian Tortoise



- Gaussian easier to estimate

- Laplacian yields more sparsity

Symmetric around μ.

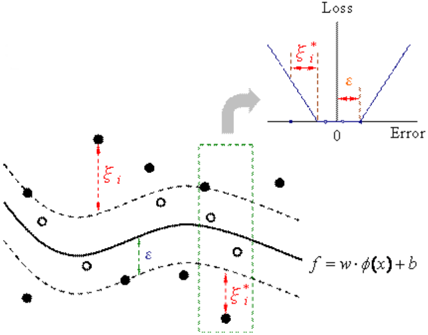
Support Vector Regression

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

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, **Support Vector Regression**
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

Support Vector Regression (SVR)



Idea: Consider only those errors that exceed a threshold

- Any point in the band (of ϵ) is not penalized. Thus the loss function is known as ϵ-*insensitive loss*
- Any point outside the band is penalized, and has slackness ξ_i or ξ_i^*
- The SVR model curve may not pass through any training point