Week 6 Revision

Regularization

Story so far....

- Probabilistic view of linear regression model:

X is the data matrix of shape $d \times n$ and $\mathbf{y}$ be the label vector of shape $n \times 1$

We assume that $y_i|\mathbf{x}_i = \mathbf{w}^T\mathbf{x}_i + \epsilon$

$$y_i|x_i \sim \text{Normal}\left(\mathbf{w}^T\mathbf{x}_i,\ \sigma^2\right)$$

$\mathbf{w}$ is a random variable??

We use MLE to estimate $\mathbf{w}$.

- How good this estimate is?
- Can we do something else to get the better estimate?

Let $\mathbf{w}_{ML}$ be the solution of the optimization problem obtained from the above MLE.

- $\mathbf{w}_{ML}$ is estimate of $\mathbf{w}$.

$$\text{Mean squared error } = E\left[||\mathbf{w} - \mathbf{w}_{ML}||^2\right]$$
$$= \sigma^2 \text{ trace}\left(\left(XX^T\right)^{-1}\right)$$

- Higher the variance in the error will confuse the estimate.
- MSE also depends upon the features.

Let $\lambda_1,\ \lambda_2,\ ...,\ \lambda_d$ are eigenvalues of XX^T, then

$$\text{MSE} = \sigma^2 \sum_{i=1}^{d} \frac{1}{\lambda_i}$$

Now, consider the following estimator:

$$\widehat{\mathbf{w}}_{new} = \left(XX^T + \lambda I\right)^{-1} Xy$$

Here $\lambda \in \mathbb{R}^+$

Then

$$\text{MSE} = \sigma^2 \sum_{i=1}^{d} \frac{1}{\lambda_i + \lambda}$$

Bayesian Modeling

- Recap: What is Bayesian estimation of a parameter θ of the distribution X??

$$y_i|\mathbf{x}_i \sim \mathsf{N}\left(\mathbf{w}^T\mathbf{x}_i,\ 1\right)$$

- Parameter: $\mathbf{w}$ of the distribution $\mathsf{N}\left(\mathbf{w}^T\mathbf{x}_i,\ 1\right)$

Let Prior for $\mathbf{w}$ be $\mathbf{w} \sim \mathsf{N}\left(0,\ \gamma^2 I\right)$, here $\gamma^2 \in \mathbb{R}^{d\times d}$ (Covariance matrix)

The sample (observations) is training dataset $\{(\mathbf{x}_1,\ y_1),\ (\mathbf{x}_2,\ y_2),\ ...,\ (\mathbf{x}_n,\ y_n)\}$

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}$$

$$P(\mathbf{w}|\ \{(\mathbf{x}_1,\ y_1),\ (\mathbf{x}_2,\ y_2),\ ...,\ (\mathbf{x}_n,\ y_n)\}) \propto P(\{(\mathbf{x}_1,\ y_1),\ (\mathbf{x}_2,\ y_2),\ ...,\ (\mathbf{x}_n,\ y_n)\}|\mathbf{w})P(\mathbf{w})$$

$$\propto \prod_{i=1}^{n} \exp\left(\frac{-\left(y_i - \mathbf{w}^T\mathbf{x}_i\right)^2}{2} - \frac{1}{2\gamma^2}||\mathbf{w}||^2\right)$$

So, the MAP estimate is

$$\widehat{\mathbf{w}}_{MAP} = \arg\max_{\mathbf{w}} \left(\frac{-\left(y_i - \mathbf{w}^T\mathbf{x}_i\right)^2}{2} - \frac{1}{2\gamma^2}||\mathbf{w}||^2\right)$$

$$= \arg\min_{\mathbf{w}} \left(\frac{\left(y_i - \mathbf{w}^T\mathbf{x}_i\right)^2}{2} + \frac{1}{2\gamma^2}||\mathbf{w}||^2\right)$$

How to solve the above optimization problem?

- Take gradient and set it to zero
- Gradient methods

Gradient $\nabla f = XX^T\mathbf{w} - X\mathbf{y} + \dfrac{\mathbf{w}}{\gamma^2}$

$$\widehat{\mathbf{w}}_{MAP} = \left(XX^T + \frac{1}{\gamma^2}I\right)^{-\dagger} X\mathbf{y}$$

MAP estimate for linear regression assuming the Gaussian prior $\mathrm{N}\left(0,\ \gamma^2 I\right)$ for $\mathbf{w}$ is same as the ridge regression.

$$L(\mathbf{w}) = \underbrace{\frac{\left(y_i - \mathbf{w}^T \mathbf{x}_i\right)^2}{2}}_{\text{Squared loss}} + \underbrace{\frac{1}{2\gamma^2}||\mathbf{w}||^2}_{\text{Regularizer}}$$

- It minimizes squared loss with penalizing the larger values of $w's$.
- The role of ridge regression is to improve the accuracy and stability of the regression model by reducing the impact of multi-collinearity.
  - multi-collinearity: a phenomenon that occurs when two or more features in a regression model are highly correlated with each other, which can lead to unstable and unreliable regression coefficients.

Why did we choose zero mean prior for $\mathbf{w}$?

- It encourages the regression coefficients to be small and centered around zero, which helps to reduce the impact of multicollinearity. When the prior mean is zero, the regularization penalty acts to shrink the coefficients towards zero, which helps to prevent them from becoming too large.
- Moreover, assuming a prior mean of zero does not introduce any bias into the model, as the mean of the posterior distribution will still be influenced by the data.
  - This means that the final estimate of the regression coefficients will be unbiased and will take into account the information from the data.

Where is MAP solution and MLE solution?

- MAP: least square Regression with regularization.
- MLE: least square Regression without regularization.

$$\arg\min_{\mathbf{w}} \left(\frac{\left(y_i - \mathbf{w}^T \mathbf{x}_i\right)^2}{2} + \lambda ||\mathbf{w}||^2 \right)$$

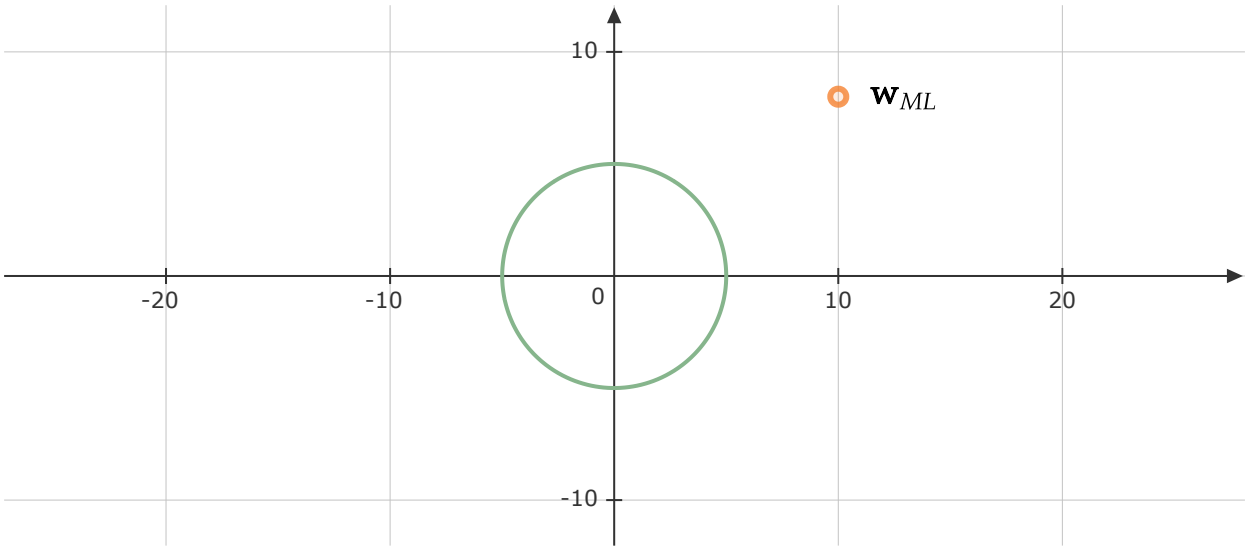
can be written as

$$\arg\min_{\mathbf{w}} \left(\frac{\left(y_i - \mathbf{w}^T \mathbf{x}_i\right)^2}{2} \right)$$

subject to

$$||\mathbf{w}||^2 \leq \theta$$

We are looking the $\mathbf{w}$ for which $||\mathbf{w}||^2 \leq \theta \Rightarrow w_1^2 + w_2^2 \leq \theta$



- $\mathbf{w}_{ML}$ incurs the minimum training loss (By design).
- We are looking for $\mathbf{w}$ which incurs loss more (say by amount c) than that of $\mathbf{w}_{ML}$.

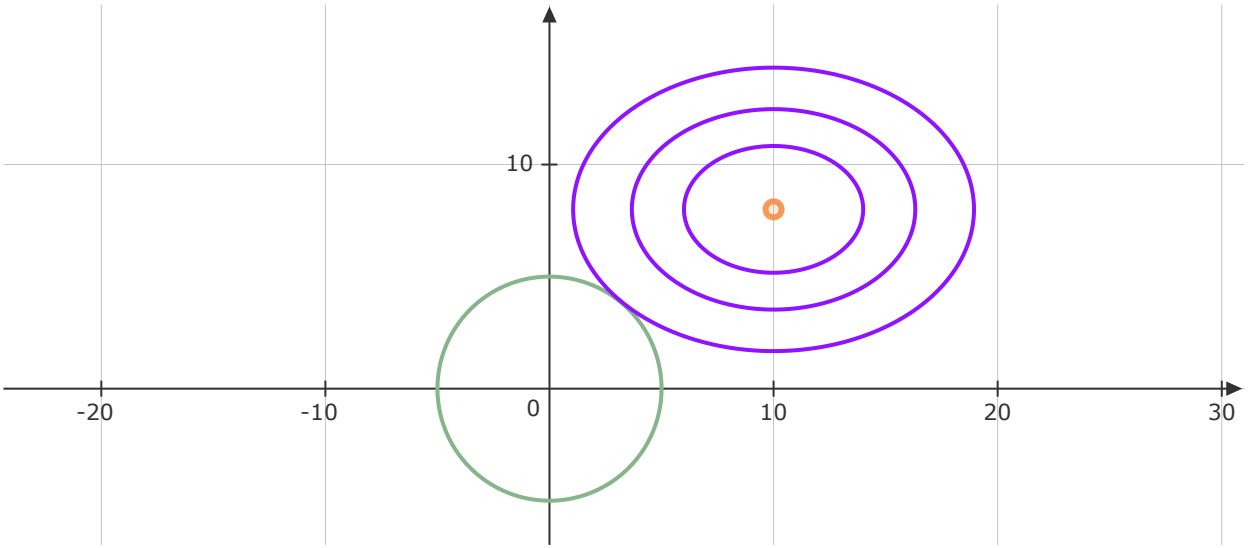
That is we are looking $\mathbf{w}$ for which

$$||X^T\mathbf{w} - \mathbf{y}||^2 \;=\; ||X^T\mathbf{w}_{ML} - \mathbf{y}||^2 + c$$

By solving this, we have

$$(\mathbf{w} - \mathbf{w}_{ML})^T\left(XX^T\right)(\mathbf{w} - \mathbf{w}_{ML}) \;=\; c'$$

This is a equation of a ellipse.


10
-20
-10
0
10
20
30

$L1$ Regularization (LASSO)

An alternate way to regularize will be to use the $L1$ norm rather than $L2$ norm.

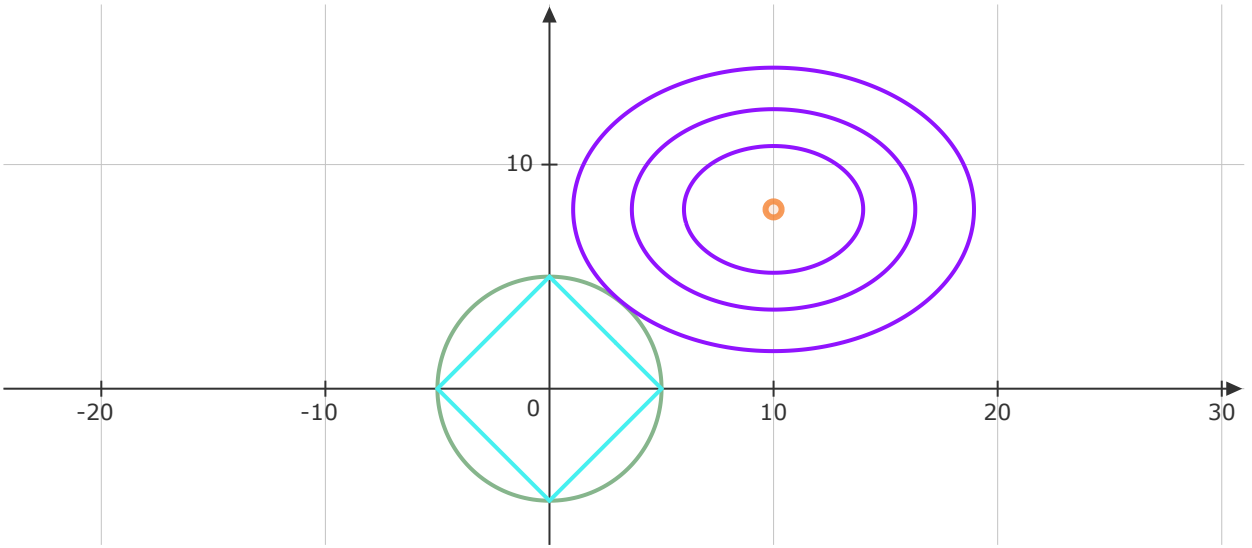
$$\arg\min_{\mathbf{w}} \left(\frac{\left(y_i - \mathbf{w}^T \mathbf{x}_i\right)^2}{2} + \lambda ||\mathbf{w}||_1 \right)$$

It can be written as

$$\arg\min_{\mathbf{w}} \left(\frac{\left(y_i - \mathbf{w}^T \mathbf{x}_i\right)^2}{2} \right)$$

subject to

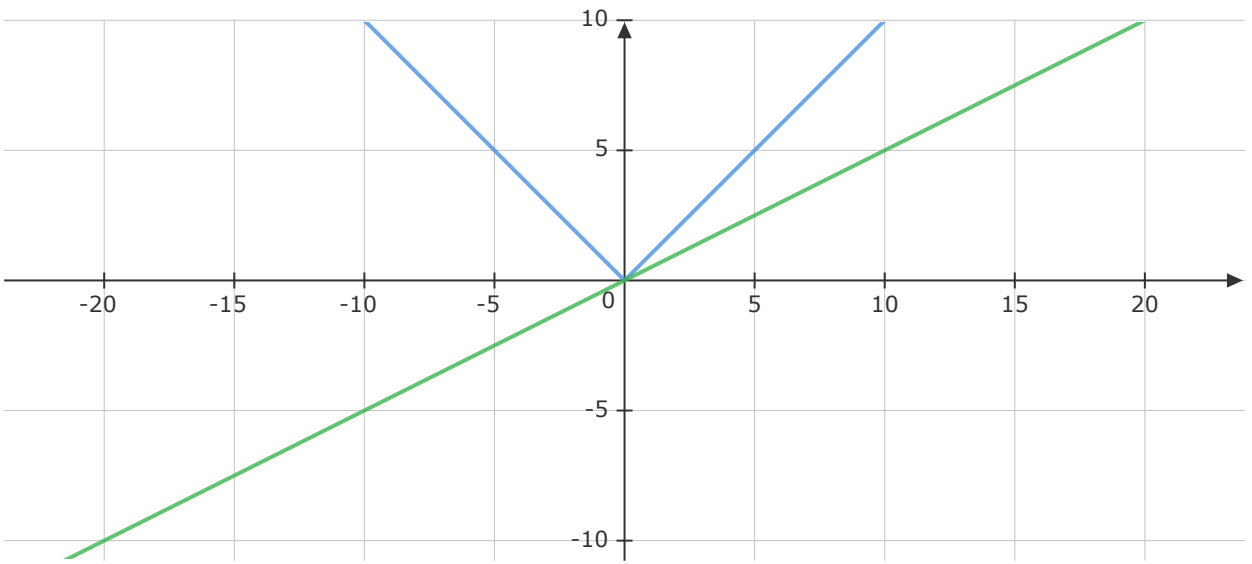
$$||\mathbf{w}||_1 \leq \theta$$


10
-20
-10
0
10
20
30

- LASSO does not have closed form solution.
- Sub-gradient methods are used to find the optima

LASSO results in more sparse weight vector.

Sub-gradient:



A vector $g \in \mathbb{R}^d$ is called sub-gradient of $f: \mathbb{R}^d \rightarrow \mathbb{R}$ at x if

$$f(z) \geq f(x) + g^T(z - x) \quad \forall z$$