Regularization

This section covers variation of linear regression that are useful when we have a large number of features.

l_{p} and l_{pq} Norms

First thing is revising the algebra for both the norms.

$$||\mathsf{X}||_{p,q} = \left(\sum_{i=1}^{n}\left(\sum_{j=1}^{m}|x_{ij}|^p\right)^{\frac{q}{p}}\right)^{\frac{1}{q}} = \left(\sum_{i=1}^{n}||\mathsf{x}^i||_p^q\right)^{\frac{1}{q}}$$

Figure 1: l_{pq} equation variation

Note that p,q $\geq$ 1, however we define the l_0-norm as the number of non-zero values in a vector. With that definition of norms, lets look at some patterns that occur at particular values for a norm. For example at the $||\mathrm{w}||_{\mathrm{p}} = 1$..

- $||\mathrm{w}||_0 = 1$, we only have one non-zero feature from the vector space
- $||\mathrm{w}||_1 = 1$, we have a diamond-like structure which hits each axis at the value of the norm, 1
- $||\mathrm{w}||_2$, we have a circular/spherical structure which hits each axis at the value of the norm, 1

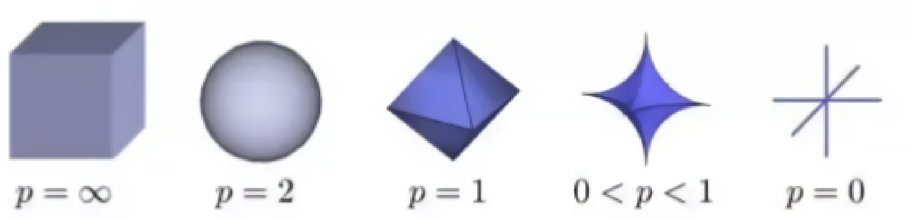

Figure 2: Shapes of $||\mathrm{w}||_{\mathrm{p}} = 1$

@@@@@@@INSERT IMAGES OF THE GRAPHS HERE@@@@@@

l_0-norm: number of nonzero values in vector

either value is on w_1 or w_2 but not both, or 0,0 neither ($|\mathrm{w}_0|$=1)

l_1-norm is $|w_1| + |w_2|$ ($|w_1|=1$)

l_2-norm is $w_1{}^2 + w_2{}^2$ ($|w_1|=2$)

Focus on 0 1 and 2 for p

Linear Regression

The objective function for linear regression, with a large number of feature, tends to want to use all the features which will result in overfitting. This is the curse of *dimensionaliy* (will be covered later).

Linear regression + regularization handles the issue of overfitting.

- Linear regression: how well it fit
- Regularization: how robust

$$\min_{\mathbf{w}} ||\mathbf{y}^T - \mathbf{w}^T \mathbf{X}||_2^2$$

Figure 3: Objective Function for Linear Regression (no regularization)

Regularization

Regularization term: placed in objective function to prevent model from overfitting. This done by adding a minimization portion to the objective function that applies to the trained coefficients of our model. The equation looks like this…

$$\min f(\mathbf{x}) + r(\mathbf{x}),$$

$f(\mathbf{x})$ is a goodness of fit function (equation in figure 3)

$r(\mathbf{x})$ is a regularization function

When minimizing a function, you can simply add a regularization term in addition, this allows you to mitigate overfitting the data.

Ridge Regression

Ridge regression model adds an l_2 regularization to reduce the values of coefficients of the model. The objective is…

$$\min_{\mathbf{w}} ||\mathbf{y}^T - \mathbf{w}^T\mathbf{X}||_2^2 + \alpha||\mathbf{w}||_2^2,$$

α is a constant hyperparameter of the model

- Adjusting α changes how sensitive the model is to the training data
- Large α: reduces how much the model focuses on data
- Small α: model focuses more on the trends in the data
- $\alpha = 0$: a model that is equivalent to the base linear regression

Ridge regression minimizes the effect of colinearity by gaining as much information as possible from the least amount of features.

- The model wont try to eliminate/zero-out a particular feature's coefficient but it will lower their values

Lasso

The lasso model adds an l_1 regularization to make some coefficients of the model go to 0. The objective is...

$$\min_{\mathbf{w}} ||\mathbf{y}^T - \mathbf{w}^T\mathbf{X}||_2^2 + \alpha||\mathbf{w}||_1$$

α is a constant hyperparameter and behaves like the ridge regression method.

Lasso regularization tries to reduce coefficients to 0, (getting rid of some of the features). Much more impact than ridge regression.

Elastic Net

Elastic net model balances between both approaches of lasso and ridge regression by utilizing both l_1 and l_2-norms. The objective for elastic net is...

$$\min_{\mathbf{w}} ||\mathbf{y}^T - \mathbf{w}^T\mathbf{X}||_2^2 + \lambda_1||\mathbf{w}||_1 + \lambda_2||\mathbf{w}||_2^2$$

λ_1 and λ_2 are the coefficients for the l_1 and l_2-norms respectively.

Can manually control these hyperparameters separately or create a relationship between the two. If creating the relationship, define two new hyperparameters, α and ρ to make a new objective function that looks like...

$$\min_{\mathsf{w}} ||\mathsf{y}^T - \mathsf{w}^T\mathsf{X}||_2^2 + \alpha\rho||\mathsf{w}||_1 + \frac{\alpha(1-\rho)}{2}||\mathsf{w}||_2^2.$$

- α is the normalization coefficient
- ρ is a balancing ratio between norms l_1 and l_2-norms

With elastic net (with either hyperparameters) the model focuses on lowering the coefficient and focuses on less features that are key to the resulting target.

If after running this model, you find that λ_1 or λ_2 is 0, than that means lasso or ridge is more optimal than the other. If they are both 0, than the resulting model is the original linear regression model.

Group Lasso

When the features of the data belong to some logical grouping, we can often incorporate the group norm.

- Group norm is calculated based on a defined grouping of the features
- Idea is to incorporate data's logical grouping when regularizing, so you can identify the most important groups of features

Group lasso is defined similar to lasso, but instead of simple l_1 regularization we use the group l_2 regularization. A group l_{p} norm is defined as…

$$||\mathsf{w}||_{g_p} = \sum_{g \in \mathsf{g}} \left\{ \sum_{j \in g} |\boldsymbol{w}_j|^p \right\}^{1/p}$$

The group l_2 norm is therefore,

$$||\mathsf{w}||_{g_2} = \sum_{g \in \mathsf{g}} \left\{ \sqrt{\sum_{j \in g} |\boldsymbol{w}_j|^2} \right\}$$

The group Lasso can be presented as

$$\min_{\mathsf{w}} ||\mathsf{y}^T - \mathsf{w}^T\mathsf{X}||_2^2 + \alpha||\mathsf{w}||_{g_2}$$

With the group lasso, the l_1 group norm induces sparsity on each group of features, attempting to eliminate any groups that don't provide as much value as others.

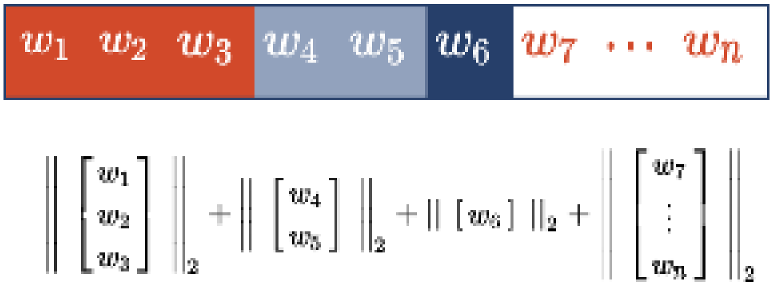
w_1 w_2 w_3 w_4 w_5 w_6 w_7 $\cdots$ w_n
$\left\| \begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix} \right\|_2 + \left\| \begin{bmatrix} w_4 \\ w_5 \end{bmatrix} \right\|_2 + \| [w_6] \|_2 + \left\| \begin{bmatrix} w_7 \\ \vdots \\ w_n \end{bmatrix} \right\|_2$

Figure 4: Example of group l_2 norm calculation for a given group of features

Sparsity Induction

With these regularization methods, we are inducing sparsity on the learned model.

- This forces the model to pick less features to use to generalize about the data.
- This generalization is what we want when adding regularization to avoid overfitting
- Another aspect of sparsity induction is to identify key features that are predictive of the target.
- By setting model to set some values to zero, we force it to learn about the actual behaviour with less features