Supervised Learning

MInDS @ Mines

In this lecture we discuss variations of linear regression that are useful when we have a large number of features. We'll cover lasso, ridge regression and elastic nets which are variations on linear regression that add regularization parameters to prevent overfitting. We want our model to generalize and adding regularization parameters is a key way to do that.

ℓ_p & $\ell_{p,q}$ norms

Before we begin discussing the common variations of linear regression that include regularizations, let's first revise some linear algebra regarding ℓ_p and $\ell_{p,q}$ norms.

An ℓ_p norm of a vector, $\mathbf{x}$ is the p^{th} root of the sum of its n values raised to the p^{th} power,

$$||\mathbf{x}||_p = \left(\sum_{i=1}^{n} x_i^p\right)^{\frac{1}{p}}. \tag{1}$$

An $\ell_{p,q}$ norm of an $n \times m$ matrix, $\mathbf{X}$, with $(\mathbf{x}_0 \dots \mathbf{x}_m)$ as its column vectors, is the ℓ_q-norm of the resulting vector from calculating the ℓ_p-norm of each row vector,

$$||\mathbf{X}||_{p,q} = \left(\sum_{i=1}^{n}\left(\sum_{j=1}^{m}|x_{ij}|^p\right)^{\frac{q}{p}}\right)^{\frac{1}{q}} = \left(\sum_{i=1}^{n}||\mathbf{x}^i||_p^q\right)^{\frac{1}{q}} \tag{2}$$

.

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, let's look at some patterns that occur at particular values for a norm, for example at the $||\mathbf{w}||_p = 1$.

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

The ℓ_1-norm of a vector is the Manhattan distance which is the total distance along each axis. This is also called the taxicab distance. The ℓ_2-norm of a vector is the Euclidean distance which is also the straight line distance between two points.

The $\ell_{p,q}$ norm of a matrix, $\mathbf{X}$ where $p = q = 2$ is called the Frobenius norm and can be represented as $||\mathbf{X}||_F$

A great post that may provide more intuition around regularizaiton is available at `https://medium.com/mlreview/l1-norm-regularization-and-sparsity-explained-for-dummies-5b0e4be3938a`.

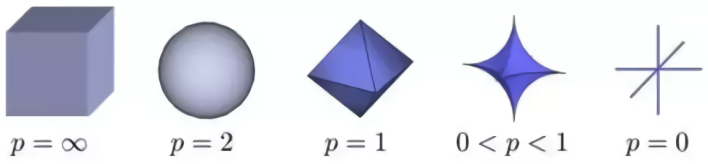

Figure 1: An overview of ℓ_p norm visualizations.

Linear Regression

We define the objective function for linear regression as the minimization of the squared Euclidean distance between the model and the true value,

$$\min_{\mathbf{w}} ||\mathbf{y}^T - \mathbf{w}^T\mathbf{X}||_2^2. \tag{3}$$

With a large number of features in $\mathbf{X}$, our model wants to use all the features and can result in overfitting. The model is also more likely to overfit when we don't have a significantly large amount of data relative to the number of features. This is as a result of the curse of dimensionality which we will cover in a later lecture when we discuss feature learning.

When the model features are highly correlated, the linear regression model is very sensitive to random noise. To handle this problem, we use regularizations.

Regularization

To prevent our model from overfitting, we can add a regularization term to the objective function. Regularization terms allow us to mitigate some of the issues that cause ordinary linear regression to overfit to our data. We do that by adding a minimization portion to the objective function that applies to the trained coefficients of our model.

In general terms, we usually solve the following function,

$$\min f(\mathbf{x}) + r(\mathbf{x}), \tag{4}$$

where $f(\mathbf{x})$ is a goodness of fit function and $r(\mathbf{x})$ is a regularization function. Generally, when minimizing a function, we can simply add a regularization term in addition to it that allows us to mitigate overfitting to the training dataset. Examples of these functions for linear regression follow.

Ridge Regression

The ridge regression model adds an ℓ_2 regularization to reduce the values of coefficients of our model. The objective is,

$$\min_{\mathbf{w}} ||\mathbf{y}^T - \mathbf{w}^T\mathbf{X}||_2^2 + \alpha||\mathbf{w}||_2^2, \tag{5}$$

where α is a constant hyperparameter of our model. We can use α to adjust how sensitive the model is to the training data. When the model is trained with a larger value for α it is reducing how the model focuses on the data. With a smaller value for α, the model focuses more on the trends in the data. When $\alpha = 0$, we end up with a model that is equivalent to the base linear regression.

When we have correlations between features and we use a regularization, the model will focus on gaining as much information as possible from

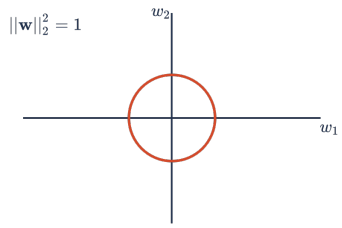

Figure 2: An example of the ℓ_2 norm set to a particular value. As we minimize this value, the circle gets a smaller radius.

the smallest number of features. This minimizes the effect of colinearity on the learned model. With ridge regression (using the ℓ_2-norm), the model won't necessarily try to completely 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 our model go to 0. The objective is,

$$\min_{\mathsf{w}} ||\mathsf{y}^T - \mathsf{w}^T\mathsf{X}||_2^2 + \alpha||\mathsf{w}||_1, \tag{6}$$

where α is a constant hyperparameter of our model. This α works similarly to the Ridge regression method's α. Just like Ridge regression, the model is trying to minimize the effect of colinearity on the learned model, however, when we use the ℓ_1-norm, the model will try to more strongly reduce the features and we will see many zero values for our coefficients.

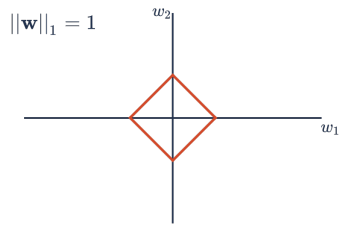

Figure 3: An example of the ℓ_1 norm set to a particular value. As we minimize this value, the diamond gets smaller.

Elastic Net

The 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 nets is,

$$\min_{\mathsf{w}} ||\mathsf{y}^T - \mathsf{w}^T\mathsf{X}||_2^2 + \lambda_1||\mathsf{w}||_1 + \lambda_2||\mathsf{w}||_2^2, \tag{7}$$

where λ_1, λ_2 are the coefficients for the ℓ_1 and ℓ_2-norms respectively. We can manually control these hyperparameters for each norm separately or we can create a relationship between λ_1 and $lambda_2$. We can define two new hyperparameters, α and ρ where α is the normalization coefficient and ρ is a balancing ratio between the ℓ_1 and ℓ_2-norms. This results in the objective as,

$$\min_{\mathsf{w}} ||\mathsf{y}^T - \mathsf{w}^T\mathsf{X}||_2^2 + \alpha\rho||\mathsf{w}||_1 + \frac{\alpha(1-\rho)}{2}||\mathsf{w}||_2^2. \tag{8}$$

With the elastic net using either hyperparameter approach, the model will focus on lowering the coefficients and focusing on less features that are key to the resulting target. Compared to Lasso, the model should have more features utilized. Compared to Ridge, the model should have more features zeroed-out. With both of these cases, if either the Lasso or Ridge regression model is a better model, when we conduct our hyperparameter search, we will find that $\lambda_1 = 0$ or $\lambda_2 = 0$. In the case where both $\lambda_1, \lambda_2 = 0$, 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 also incorporate the group norm. The group norm is calculated based on a defined grouping of the features. The idea here is to incorporate the

data's logical grouping when regularizing so that you can identify the most important groups of features.

Group Lasso is defined similar to Lasso but Instead of the simple ℓ_1 regularization we use the group ℓ_2 regularization. A group ℓ_p norm is defined as,

$$||\mathsf{w}||_{g_p} = \sum_{g \in \mathsf{g}} \left\{ \sum_{j \in g} |\boldsymbol{w}_j|^p \right\}^{1/p} \tag{9}$$

The group ℓ_2 norm is therefore,

$$||\mathsf{w}||_{g_2} = \sum_{g \in \mathsf{g}} \left\{ \sqrt{\sum_{j \in g} |\boldsymbol{w}_j|^2} \right\} \tag{10}$$

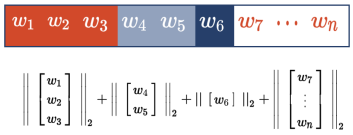

Figure 4: An example of the group ℓ_2 norm calculation for a given grouping of features.

The group Lasso can be presented as,

$$\min_{\mathsf{w}} ||\mathsf{y}^T - \mathsf{w}^T\mathsf{X}||_2^2 + \alpha||\mathsf{w}||_{g_2}, \tag{11}$$

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.

Sparsity Induction

With these regularization methods, we are *inducing sparsity* on the learned model. We are essentially forcing the model to pick less features to use to generalize about the data. This generalization is what we're after when the focus is to add regularization to prevent overfitting. Another aspect of sparsity induction, however, is the ability to identify key features that are predictive of the target. By telling the model to set some values to zero, we force it to learn the select few features that determine the large share of actual behavior. This leads to the model ignoring noise in the data and therefore avoiding overfitting. It also leads to the model being interpretable and simple since we can focus on a select set of features and understand how they are predictive of the target.

A dataset, or matrix is said to be sparse if it has many zero values.