AIFS ML Lecture 7: Machine Learning Basics

Suraj Narayanan Sasikumar

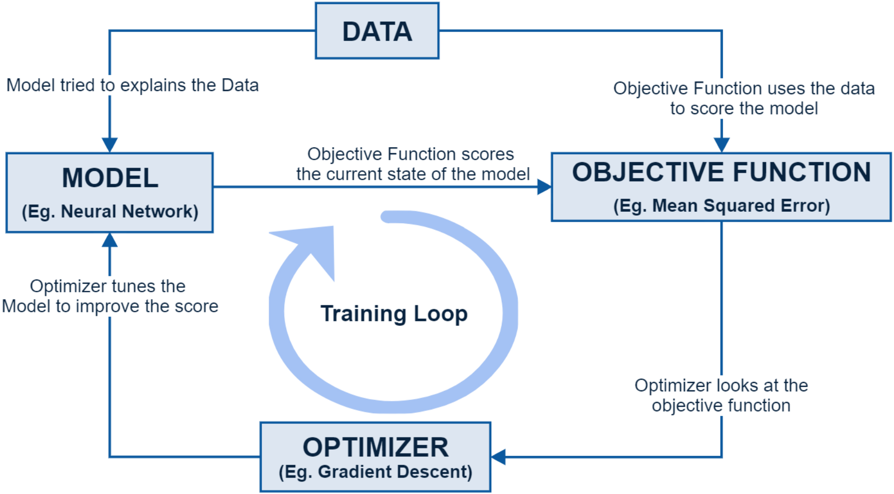
Hessian AI Labs

May 13, 2020

Overview

Table of Contents

Recap - Learning Algorithm



Recap

- When a model is overfitting, we observed that the values of the model parameters are large and alternating.
- *Regularization* is the process of adding a penalty term to the loss-function of a learning algorithm to prevent overfitting.
- Based on the observation, the penalty should be the length of the parameter vector.
- *Norm* of a vector is a measure of its length in relation to the $\mathbf{0}$ vector (origin).
- Two commonly used norms are the ℓ_1 and ℓ_2 norm.

Recap - LASSO Regression

- When the ℓ_1-norm of the parameter vector ($\mathbf{w}$) is added to the loss function of linear regression, the resulting algorithm is called *LASSO Regression*.
- Encourages sparsity in parameter vector ($\mathbf{w}$)
- Components of the algorithm
  - **Data:** $\boldsymbol{X}, \mathbf{y}$
  - **Model:** $\hat{y} = \mathbf{w}^T\mathbf{x}$ (y-intercept included in the dot-product, $x_0 = 1$)
  - **Loss Function:** Minimize Mean Squared Error with ℓ_1 penalty

$$\text{MSE}(\mathbf{w}) = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{w}^T\mathbf{x} - y_i)^2 + \lambda||\mathbf{w}||_1$$

  - **Optimizer:** LASSO Regression does not have a closed form solution. For LASSO, the following two iterative optimization algorithms are used:
    1. LARS: least-angle regression.
    2. Coordinate Descent. Minimize over one dimension (coordinate) at a time.

Recap - Ridge Regression

- When the squared ℓ_2-norm of the parameter vector ($\mathbf{w}$) is added to the loss function of linear regression, the resulting algorithm is called *Ridge Regression*.
- Performs better when many features are known to be correlated.
- Components of the algorithm
  - **Data:** $\boldsymbol{X}, \mathbf{y}$
  - **Model:** $\hat{y} = \mathbf{w}^T\mathbf{x}$ (y-intercept included in the dot-product, $x_0 = 1$)
  - **Loss Function:** Minimize Mean Squared Error with ℓ_2 penalty

$$\text{MSE}(\mathbf{w}) = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{w}^T\mathbf{x} - y_i)^2 + \lambda||\mathbf{w}||_2^2$$

  - **Optimizer:** Closed Form Solution obtained by setting gradient of MSE to 0, $\nabla_{\mathbf{w}}\text{MSE}(\mathbf{w}) = 0$

$$\mathbf{w}^\star = (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\mathbf{y}$$

Table of Contents

Regularization Coefficient

- What should the value of λ be in LASSO and Ridge Regression?

Regularization Coefficient

- What should the value of λ be in LASSO and Ridge Regression?
- λ is a hyperparameter like M in polynomial regression. It is an algorithm-hyperparameter since it's a parameter of the loss function.

Regularization Coefficient

- What should the value of λ be in LASSO and Ridge Regression?
- λ is a hyperparameter like M in polynomial regression. It is an algorithm-hyperparameter since it's a parameter of the loss function.
- How does the fit of the model change based on the value of λ?

Regularization Coefficient

- What should the value of λ be in LASSO and Ridge Regression?
- λ is a hyperparameter like M in polynomial regression. It is an algorithm-hyperparameter since it's a parameter of the loss function.
- How does the fit of the model change based on the value of λ?
- A larger value for λ means $||\mathbf{w}||$ should be small. Hence the capacity of the model is decreased as the ability to "tune" the model is constrained.

Regularization Coefficient

- What should the value of λ be in LASSO and Ridge Regression?
- λ is a hyperparameter like M in polynomial regression. It is an algorithm-hyperparameter since it's a parameter of the loss function.
- How does the fit of the model change based on the value of λ?
- A larger value for λ means $||\mathbf{w}||$ should be small. Hence the capacity of the model is decreased as the ability to "tune" the model is constrained.
- Very large λ value can lead to *Underfitting*.

Regularization Coefficient

- What should the value of λ be in LASSO and Ridge Regression?
- λ is a hyperparameter like M in polynomial regression. It is an algorithm-hyperparameter since it's a parameter of the loss function.
- How does the fit of the model change based on the value of λ?
- A larger value for λ means $||\mathbf{w}||$ should be small. Hence the capacity of the model is decreased as the ability to "tune" the model is constrained.
- Very large λ value can lead to *Underfitting*.
- A very small value for λ means a large value for $||\mathbf{w}||$ is allowed. Hence the capacity of the model is increased as more $\mathbf{w}$ values are allowed.

Regularization Coefficient

- What should the value of λ be in LASSO and Ridge Regression?
- λ is a hyperparameter like M in polynomial regression. It is an algorithm-hyperparameter since it's a parameter of the loss function.
- How does the fit of the model change based on the value of λ?
- A larger value for λ means $||\mathbf{w}||$ should be small. Hence the capacity of the model is decreased as the ability to "tune" the model is constrained.
- Very large λ value can lead to *Underfitting*.
- A very small value for λ means a large value for $||\mathbf{w}||$ is allowed. Hence the capacity of the model is increased as more $\mathbf{w}$ values are allowed.
- Very small λ values can lead to *Overfitting*. $\lambda = 0$ is the same as Linear Regression.

Table of Contents

Optimization vs Learning

- The goal of an optimization algorithm is to choose a set of inputs that minimizes an objective function.
- The goal of a machine learning algorithm is not to minimize the objective function, but rather to learn about the underlying data-generating process so that it can perform tasks like classification, regression, clustering etc.
- Overfitting is what happens when we treat a machine learning algorithm as an optimization algorithm; by explaining the noise, it perfectly models all the data points and achieves zero mean-squared error.
- The key difference is that a properly trained machine learning algorithm is able to **generalize** to data points not seen before, ie it has predictive power.