CS M146 Quiz Week 5 Solutions:

February 23, 2023

Overfitting and Underfitting

- 1.1 FALSE
- 1.2 TRUE
- 1.3 Logistic Regression: Add L2 regularization on the weight parameters.

Overfitting occurs when you achieve a good fit of your model on the training data, while it does not generalize well on new, unseen data(test dataset). Underfitting occurs when you have a model that can neither model the training data nor generalize to unseen data. Hence, if a model is **overfitting**, then adding additional training examples can improve the model performance on unseen data (decreasing the test error) while increasing the training error. On the other hand, if a model is **underfitted**, then adding new training examples does not help and we need to apply alternate machine-learning algorithms to our data.

Ridge regression

- 2.1 TRUE
- 2.2 TRUE
- 2.3 FALSE
- 2.4 FALSE

For every $n \times m$ matrix X and $\lambda > 0$, $XX^T + \lambda I$ is invertible. Therefore, ridge regression has always a unique solution defined as : $\hat{\beta}_{ridge} = (XX^T + \lambda I)^{-1}X^T y$, for $\lambda > 0$.
When $\lambda = 0$, then ridge regression estimator $\hat{\beta}_{ridge}$ is same as linear regression estimator (or OLS estimator), $\hat{\beta}_{ols} = (XX^T)^{-1}X^T y$. OLS estimator has a unique solution if (XX^T) is invertible. $X^T X$ is invertible if and only if the columns of X are linearly independent.

For a given $n \times m$ matrix X:

- If $n < m$, then its columns are always linearly dependent (m vectors in n-dimensional space are always linearly dependent for $n < m$), so $X^T X$ is not invertible.
- If $n \geq m$, then its columns either linearly dependent or linearly independent. ($n \geq m$ is a necessary condition to have a $n \times m$ matrix with linearly independent columns but it is not sufficient.)

Ridge vs linear regression estimators

- 3.1 $\frac{\sum_i x_i y_i}{\sum_i x_i^2 + \lambda}$
- 3.2 $|\hat{\beta}_{ridge}| \leq |\hat{\beta}_{linear}|$

To compute the ridge regression estimator or linear regression estimator, we need to compute the derivate of the corresponding cost function with respect to β and set the derivative to zero.

- $\hat{\beta}_{ridge} = \frac{\sum_i x_i y_i}{\sum_i x_i^2 + \lambda}$
- $\hat{\beta}_{linear} = \frac{\sum_i x_i y_i}{\sum_i x_i^2}$

Regularization on Logistic Regression

- 4.1 $\{\beta_0 = 1, \beta_1 = 1, \beta_2 = -1\}$. Lines in other options cannot either separate the dataset or correctly separate the data points with correct labels (wrong w direction).
- 4.2 If we use a very large value of λ and $j = 1$ (which will enforce $\beta_1 = 0$), we can still learn an LR classifier that achieves zero training error on this dataset. $\beta_1 = 0$ will let the LR model only classify based on x_2, which is possible to perfectly classify the data such as $-x_2 + 1 = 0$. However, LR models that use x_1 feature ($\beta_2 = 0$) or pass through the origin ($\beta_0 = 0$) cannot perfectly classify the data.
- 4.3 No. Adding regularization on bias terms will limit model capacity. As shown in the example above, regularization on bias terms tends to enforce the model passing the origin and penalize a large value of β_0.

Transformation of input features

Yes, the transformed data points are linearly separable.