Reducing Bias via High Dimensional Feature Interactions

Recall that the bias component of the test error is a result of an incorrect assumption: you may be training a linear classifier when the labels can only be explained with a non-linear classifier.

You can reduce the bias of models on such datasets by transforming the input feature vectors to higher dimensional feature vectors. Formally, for a data vector $\mathbf{x} \in \mathbb{R}^d$, you apply a transformation $\mathbf{x} \rightarrow \phi(\mathbf{x})$, where $\phi(\mathbf{x}) \in \mathbb{R}^D$ is the new feature vector. Usually $D \gg d$ because we add dimensions that capture non-linear interactions among the original features, as in the example of circles you saw in the preceding video.

In an extreme case, you could consider expanding $\mathbf{x} \in \mathbb{R}^d$ with all possible multiplicative interactions between any two dimensions. That is,

$$\text{Assume } \mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{pmatrix}, \text{ define } \phi(\mathbf{x}) = \begin{pmatrix} 1 \\ x_1 \\ \vdots \\ x_d \\ x_1x_2 \\ \vdots \\ x_{d-1}x_d \\ x_1x_2x_3 \\ \vdots \\ x_1x_2\cdots x_d \end{pmatrix}.$$

This new representation, $\phi(\mathbf{x})$, is very expressive and allows for complicated non-linear decision boundaries. However, the dimensionality is extremely high (2^d). The high dimensionality is problematic because, if we were training a linear SVM on $\phi(\mathbf{x})$ as feature vectors for example, then every dot product between $\mathbf{w}$ and $\phi(\mathbf{x})$ would take on the order of 2^d for element-wise multiplication and on the order of 2^d for summation (thus on the order of 2^{d+1} operations). This exponential increase in dimensionality would prohibitively slow our algorithm.

To summarize, the following is true about the projection into a new feature space:

Advantage: It is simple, and your problem stays convex and well behaved. Thus, you can still use gradient descent to solve the optimization problem, just in a higher higher dimensional vector space.

Disadvantage: $\phi(\mathbf{x})$ might be too high dimensional, which could slow down our training algorithm.

◂ Previous

Next ▸