Approximation-Generalization Tradeoff: VC dimension and Bias-variance trade-off

Mar 15 ,2019

XueYu

Abstract

In this post,I will write down something about Approximation-Generalization Tradeoff.Mainly introduce VC dimension and Bias-variance trade-off.

Reference

Vapnik–Chervonenkis dimension
Bias-variance tradeoff
CS229 Bias-Variance and Error Analysis, Yoann Le Calonnec, October 2, 2017
Foundations of Statistical Learning,Jiaming Mao,Xiamen University

VC dimension

➢ definition

In Vapnik–Chervonenkis theory,the Vapnik–Chervonenkis (VC) dimension is a measure of the capacity (complexity, expressive power, richness, or flexibility) of a space of functions that can be learned by a statistical classification algorithm. It is defined as the cardinalityof the largest set of points that the algorithm can shatter.

➢ VC dimension of a set-family

The growth function for a hypothesis set H, denoted m_H (N), is the maximum possible number of dichotomies H can generate on a data set of N points.
If H is capable of generating all possible dichotomies on x1, . . . , xN,then H shatters x1, . . . , xN, in which case m_H (N) = 2^N.
The Vapnik-Chervonenkis (VC) dimension of H, denoted dvc (H),is the size of the largest data set that H can shatter. dvc (H) is the largest value of N for which m_H (N) = 2^N.

➢ In statistical learning theory

The VC dimension can predict a probabilistic upper bound on the test error of a classification model. Vapnik proved that the probability of the test error distancing from an upper bound (on data that is drawn i.i.d.from the same distribution as the training set) is given by:

$$\Pr\left(\text{test error} \leqslant \text{training error} + \sqrt{\frac{1}{N}\left[D\left(\log\left(\frac{2N}{D}\right)+1\right)-\log\left(\frac{\eta}{4}\right)\right]}\right) = 1-\eta,$$

Where D is the VC dimension of the classification model, 0≤η≤1, and N is the size of the training set (restriction: this formula is valid when D 《 N. When D is larger, the test-error may be much higher than the training-error. This is due to overfitting).
The VC dimension also appears in sample-complexity bounds.

Bias-variance trade-off

Bias-variance decomposition provides another way of looking at the approximation-generalization tradeoff. In statistics and machine learning, the bias–variance tradeoff is the property of a set of predictive models whereby models with a lower bias in parameter estimation have a higher variance of the parameter estimates across samples, and vice versa.

- Underfitting

High bias can cause an algorithm to miss the relevant relations between features and target outputs. In other words,a model does not fit the data well enough,then we call this underfitting.

- Overfitting

High variance can cause an algorithm to model the random noise in the training data, rather than the intended outputs. In other words,a model fits its data too well, we call this overfitting.

So,people want to choose a model that both accurately captures the regularities in its training data, but also generalizes well to unseen data,which is typically impossible to do both simultaneously. High-variance may be able to represent their training set well but are at risk of overfitting to noisy or unrepresentative training data. In contrast,if we use simple models then we may get low variance that don't tend to overfit but may underfit their training data, failing to capture important regularities.
To understand the tradeoff between bias and variance, I will start by looking at the mean squared error (MSE):

$$E\left(y_0 - \hat{f}(x_0)\right)^2 = \operatorname{Var}(\hat{f}(x_0)) + [\operatorname{Bias}(\hat{f}(x_0))]^2 + \operatorname{Var}(\epsilon).$$

From equation (1),the expected mean squared error can be broken down into three sections.:the variance of the model, the squared bias of the model and the variance of the error term.
Intuitively, to get low error, we need low bias and low variance. If we increase samples size the variance term will decreace.When we choose a simple model maybe we will decrease the variance term but we can’t capture some pattern, the bias is high, and we are under-fitting.According to slide writing by Jiaming Mao,I learn that if we have a large number of datasets then we can choose a more complex model.This can connect with the knowledge of VC dimension.

Conclusion

➢ VC dimension

If the model is more complex(notice:H is more complex and dvc will increase) then we will have a better chance of approximating f in sample($E_{in} \approx 0$).On the other hand, If the model is less complex(notice: :H is less complex and dvc will decrease)then it is a better chance of generalizting out of sample($E_{in} \approx E_{out}$).Otherwise,when we have a large number of datasets we can choose a more complex model.

➢ Bias-variance trade-off

When the model is more complex,the squared bias of the model will turn smaller but the variance of the model may be larger than before.The same as above when we have a large number of datasets we can choose a more complex model.