Lecture 5: Challenges to Machine Learning

DD2431

Atsuto Maki

Autumn, 2014

Overfitting

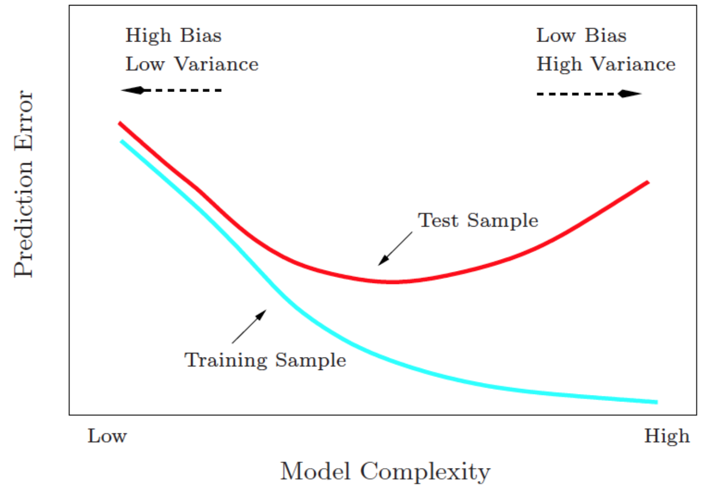
Visited in Lecture 2 using decision tree.

Good results on training data, but generalizes poorly.
This occurs due to

- Non-representative sample
- Noisy examples

Overfitting

When the learned models are overly specialized for the training samples.



(T. Hastie et al, The Elements of Statistical Learning)

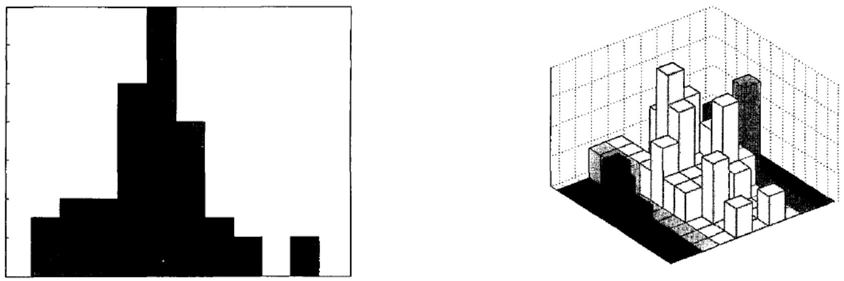
Curse of Dimensionality

Curse of Dimensionality

Imagine: inputs represented by 30 features but some of them are less relevant to target function. Will you use all of them?

- Easy problems in low-dimensions are harder in high-dimensions
  – training more complex model with limited sample data
- In high-dimensions everything is far from everthing else
  – issues in Nearest Neighbours

Example: Normal random numbers in 1-d and 2-d
(both plots for 100 inputs)



Too few data to represent the probability density function in 2-d.

Intuitions in low-dimensions do not apply in high-dimensions
Real world is in 3-d, but we deal with data for instance in 1000-d

- Uniform distribution on hypercube
- Volume of hypersphere

Techniques for dimensionality reduction / feature selection exist.

The Bias-Variance Trade-off

Concepts of prediction errors

Let us imagine we could repeat the modeling for many times – each time by gathering new set of training samples, $\mathcal{D}$.

The resulting models will have a range of predictions due to randomness in the underlying data set.

- Error due to **Bias**: the difference between the average (expected) prediction of our model and the correct value.
- Error due to **Variance**: the variability of a model prediction for a given data point between different realizations of the model.

The bias-variance decomposition

Let us consider

$$\begin{aligned} f(\mathbf{x}) &: \text{true function} \\ \hat{f}(\mathbf{x}) &: \text{prediction function (= model) estimated with } \mathcal{D} \\ E[\hat{f}(\mathbf{x})] &: \text{average of models due to different sample sets} \end{aligned}$$

The mean square error (MSE) for estimating $f(\mathbf{x})$

$$\begin{aligned} E_{\mathcal{D}}[(\hat{f}(\mathbf{x}) - f(\mathbf{x}))^2] &= E_{\mathcal{D}}[(\hat{f}(\mathbf{x}) - E[\hat{f}(\mathbf{x})])^2] + (E[\hat{f}(\mathbf{x})] - f(\mathbf{x}))^2 \\ &= \textbf{Variance} + (\textbf{Bias})^2 \end{aligned}$$

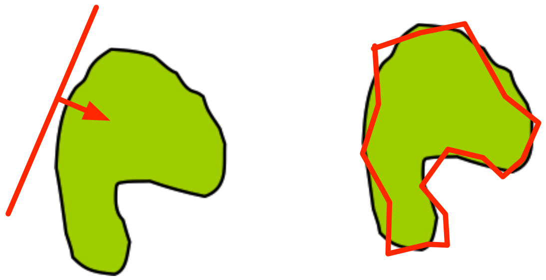
Bias of a classifier is the discrepancy between its averaged estimated and true function

$$E[\hat{f}(\mathbf{x})] - f(\mathbf{x})$$

(derivation of decomposition at the lecture)

Characterization of a classifier: Bias

Green region is the true boundary.



High-bias classifier **Low-bias classifier**

Low model complexity (small # of d.o.f.) $\Longrightarrow$ High-bias
High model complexity (large # of d.o.f.) $\Longrightarrow$ Low-bias

Characterization of a classifier: Variance

Variance of a classifier is the expected divergence of the estimated prediction function from its average value:

$$E_{\mathcal{D}}[(\hat{f}(\mathbf{x}) - E[\hat{f}(\mathbf{x})])^2]$$

This measures how dependent the classifier is on the random sampling made in the training set.

Low model complexity (small # of d.o.f.) $\Longrightarrow$ Low-variance
High model complexity (large # of d.o.f.) $\Longrightarrow$ High-variance

decision trees

Low bias classifiers produce decision boundaries which on average are good approximations to the true decision boundary.

High variance classifiers produce differing decision boundaries which are highly dependent on the training data.

Our intuition may tell:

- The presence of bias indicates something basically wrong with the model and algorithm...
- Variance is also bad, but a model with high variance could at least predict well on average...

So the model should minimize bias even at the expense of variance??

Not really!

Bias and variance are equally important as we are always dealing with a single realization of the data set.