Learning Theory Part 3: Bias-Variance Tradeoff

CS 760@UW-Madison
Goals for the lecture

you should understand the following concepts
  • estimating bias and variance
  • the bias-variance decomposition
Estimation bias and variance

- How will predictive accuracy (error) change as we vary $k$ in $k$-NN?

- Or as we vary the complexity of our decision trees?

- The bias/variance decomposition of error can lend some insight into these questions.

Note that this is a different sense of bias than in the term *inductive bias*. 
Background: Expected values

• the expected value of a random variable that takes on numerical values is defined as:

\[ E[X] = \sum_x x P(x) \]

this is the same thing as the mean

• we can also talk about the expected value of a function of a random variable

\[ E[g(X)] = \sum_x g(x) P(x) \]
Defining bias and variance

• consider the task of learning a regression model \( f(x; D) \) given a training set \( D = \{(x^{(1)}, y^{(1)}),..., (x^{(m)}, y^{(m)})\} \)

• a natural measure of the error of \( f \) is

\[
E[(y - f(x; D))^2 | D]
\]

where the expectation is taken with respect to the real-world distribution of instances
Defining bias and variance

- this can be rewritten as:

\[
E[(y - f(x; D))^2 | x, D] = E[(y - E[y|x])^2 | x, D] + (f(x; D) - E[y|x])^2
\]

- error of \( f \) as a predictor of \( y \)
- noise: variance of \( y \) given \( x \); doesn’t depend on \( D \) or \( f \)
Defining bias and variance

- now consider the expectation (over different data sets $D$) for the second term

$$E_D[(f(x; D) - E[y | x])^2] =$$

$$\left( E_D[f(x; D)] - E[y | x] \right)^2$$  \text{bias}

$$+ E_D[(f(x; D) - E_D[f(x; D)])^2]$$  \text{variance}

- bias: if on average $f(x; D)$ differs from $E[y | x]$ then $f(x; D)$ is a biased estimator of $E[y | x]$

- variance: $f(x; D)$ may be sensitive to $D$ and vary a lot from its expected value
Bias/variance for polynomial interpolation

- the 1\textsuperscript{st} order polynomial has high bias, low variance
- 50\textsuperscript{th} order polynomial has low bias, high variance
- 4\textsuperscript{th} order polynomial represents a good trade-off
Bias/variance trade-off for k-NN regression

- consider using $k$-NN regression to learn a model of this surface in a 2-dimensional feature space
Bias/variance trade-off for k-NN regression

- Bias for 1-NN
- Variance for 1-NN
- Bias for 10-NN
- Variance for 10-NN

darker pixels correspond to higher values
Bias/variance trade-off

- consider $k$-NN applied to digit recognition
Bias/variance discussion

• predictive error has two controllable components
  • expressive/flexible learners reduce bias, but increase variance

• for many learners we can trade-off these two components (e.g. via our selection of $k$ in $k$-NN)

• the optimal point in this trade-off depends on the particular problem domain and training set size

• this is not necessarily a strict trade-off; e.g. with ensembles we can often reduce bias and/or variance without increasing the other term
Bias/variance discussion

the bias/variance analysis
  • helps explain why simple learners can outperform more complex ones
  • helps understand and avoid overfitting
THANK YOU

Some of the slides in these lectures have been adapted/borrowed from materials developed by Mark Craven, David Page, Jude Shavlik, Tom Mitchell, Nina Balcan, Elad Hazan, Tom Dietterich, and Pedro Domingos.