CSE 575: Statistical Machine Learning

Jingrui He

CIDSE, ASU

Bias-Variance Tradeoff

Bias–Variance Decomposition of Error

- Consider simple regression problem f:X→T

$$t = f(x) = \underbrace{g(x)}_{\text{deterministic}} + \underbrace{\varepsilon}_{\text{noise} \sim N(0,\sigma)}$$

Collect some data, and learn a function h(x)

What are sources of prediction error?

Bias-Variance Tradeoff – Intuition

- Model too “simple” ! does not fit the data well
  - A biased solution

- Model too complex! small changes to the data, solution changes a lot
  - A high-variance solution

Assume target function: $t = f(x) = g(x) + \varepsilon$

(Squared) Bias of learner

- Given dataset *D* with *m* samples, learn function h(x)
- If you sample a different datasets, you will learn different h(x)
- **Expected hypothesis**: $E_D[h(x)]$

- **Bias:** difference between what you expect to learn and truth
  - Measures how well you expect to represent true solution
  - Decreases with more complex model

$$bias^2 = \int_x \{E_D[h(x)] - g(x)\}^2 p(x)dx$$

Assume target function: $t = f(x) = g(x) + \varepsilon$

Variance of learner

- Given a dataset *D* with *m* samples, you learn function h(x)
- If you sample a different datasets, you will learn different h(x)
- **Variance:** difference between what you expect to learn and what you learn from a particular dataset
  - Measures how sensitive learner is to specific dataset
  - Decreases with simpler model

$$\begin{aligned} \bar{h}(x) &= E_D[h(x)] \\ variance &= \int E_D[(h(x) - \bar{h}(x))^2] p(x) dx \end{aligned}$$

Assume target function: $t = f(x) = g(x) + \varepsilon$

Bias-Variance Tradeoff

- Choice of hypothesis class introduces learning bias
  - More complex class → less bias
  - More complex class → more variance

Assume target function: $t = f(x) = g(x) + \varepsilon$

Sources of Error 1 – Noise

- What if we have perfect learner, infinite data?
  - If our learning solution h(x) satisfies h(x)=g(x)
  - Still have remaining, *<u>unavoidable error</u>* of σ^2 due to noise ε

$$error(h) = \int_x \int_t (h(x) - t)^2 p(f(x) = t|x) p(x) dt dx$$

Assume target function: $t = f(x) = g(x) + \varepsilon$

Sources of Error 2 – Finite Data

- What if we have imperfect learner, or only m training examples?
- What is our expected squared error per example?
  - Expectation taken over random training sets *D* of size m, drawn from distribution P(X,T)

$$E_D\left[\int_x \int_t \{h(x) - t\}^2 p(f(x) = t|x)p(x)dtdx\right]$$

Assume target function: $t = f(x) = g(x) + \varepsilon$

Bias–Variance Decomposition of Error

Then expected sq error over fixed size training sets *D* drawn from P(X,T) can be expressed as sum of three components:

$$E_D\left[\int_x \int_t (h(x)-t)^2 p(t|x)p(x)dtdx\right]$$
$$= unavoidableError + bias^2 + variance$$

Where:

$$unavoidableError = \sigma^2$$
$$bias^2 = \int (E_D[h(x)] - g(x))^2 p(x)dx$$
$$\bar{h}(x) = E_D[h(x)]$$
$$variance = \int E_D[(h(x) - \bar{h}(x))^2]p(x)dx$$