Bias Variance Decomposition

23 September 2024 18:40

Empirical Risk Minimization.

Available dataset (X,Y) = collection of samples from $P(X,Y)$ ← Underlying joint probability distribution unknown to us.

↓

Goal of model development — iteratively train f_w using (X,Y)

↓

identify optimal function f_w^* from all possible hypothesis f_w.

Optimality obtained when model trained on $(X,Y)^{train}$ results in minimal cost $Q(w)$ on $(X,Y)^{test}$ → Proxy to future unknown data.

$Q(w) = L(Y, f_w(X))$ → cost is a function of random input data pair (X,Y)

One realization for $P(X,Y)$ — Measures generalization error of model.

we want model that produces minimal expected cost across all variations of $(X,Y)^{test}$

RISK

↓

Risk for model f_w:

$$R(f_w) = E[Q(w)] = E[L(Y, f_w(X))] = \int L(Y, f_w(X))\, dp(X,Y)$$

Integral over all possible realizations of data based on joint distribution $P(X,Y)$

Realization = observed value of random variable.

↓

for regression:

$$R(f_w) = E[y - f_w(x)]^2$$

Since joint distribution $P(X,Y)$ is unknown, true risk approx. via empirical risk

↓

MSE on $(X,Y)^{test}$.

$$R_{empirical}(f_w) = \frac{1}{n}\sum_{i=1}^{n}(y_i - f_w(x_i))^2$$

n = no. of samples; y_i = true value; $f_w(x_i)$ = output of model.

↓

$$f_w^* = \underset{f_w \in F_w}{\operatorname{argmin}}\, R_{empirical}(f_w)$$

f_w^* = one that minimizes empirical risk based on test data.

In linear regression: f is a linear model parameterized by w.

↳ If real mapping function is g: $Y = g(X) + \epsilon$ ← additive random noise ↓ Gaussian ($\mu = 0$, $\sigma = \sigma$)

⇒ We can build a model such that $E[f(X)] = E[g(X) + \epsilon \mid x = X]$

Now, bias = lack of capacity in model to perfectly fit g (underfitting)

variance = error due to fitting a spurious relation by random noise (overfitting)

↳ seem related, but really not.

Consider an unseen case X_{new}:

Let — $f(X_{new}) = f$ (APPROX)

$g(X_{new}) = g$ (TRUE) → deterministic + non-random.

$$R(f) = E[(y - f(x_{new}))^2 \mid x = x_{new}]$$

$$= E[(y-f)^2]$$

$$= E[(y-g+g-f)^2]$$

$$= E[(y-g)^2] + E[(g-f)^2] + 2E[(y-g)(g-f)]$$

$$= E[(y-g)^2] + E[(g-f)^2] + 2E[(y-g)(g-f)]$$

$y = g + \varepsilon$

math later

$$= E[\varepsilon^2] + E(g-f)^2$$

$$= \sigma^2 + E[g - E(f)]^2 + E[f - E(f)]^2$$

bias

variance.

$$R(f) = \sigma^2 + \text{Bias}^2(f) + \text{Var}(f)$$

σ^2: cannot be reduced

$\text{Bias}^2(f) + \text{Var}(f)$: controllable by training

Simple model
- high bias
- low variance

Complex model
- low bias
- high variance.