Statistical Justifications for regression

- Sample pts from unknown dist d.
- y-values are sum of unknown, non-random fn + random noise
- $\forall x_i$, $y_i = \underbrace{g(x_i)}_{\text{true}} + \underbrace{\varepsilon_i}_{\text{noise}}$ $\quad \varepsilon_i \sim D'$, D' has $\mu = 0$
- goal: find h that estimates g.
- choose $h(x) = \underbrace{\mathbb{E}_y[y \mid x = x]}_{\text{hope this exists}} = g(x) + \mathbb{E}[\varepsilon] = g(x)$

Least Squares → follows max likelihood

Suppose $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$; $y_i \sim \mathcal{N}(g(x_i), \sigma^2)$

$$\log f(y_i) = \frac{-(y_i - \mu)^2}{2\sigma^2} - \text{constant} \quad \Leftarrow \mu = g(x_i)$$

$$l(g; X, y) = \ln f(y_1) \cdots f(y_n) = \sum_{i=1}^{n} f(y_i)$$

$$= -\frac{1}{2\sigma^2} \boxed{\sum_{i=1}^{n} (y_i - g(x_i))^2} - \text{constant}$$

max likelihood on g is equivalent to g from least squares regression

✱ least squares works well for normal dist. noise.

Empirical Risk

- risk is expected loss $R(h) = \mathbb{E}[L]$ over all $x \in \mathbb{R}^d$ $y \in \mathbb{R}$
- discriminative model -> don't know X's dist
- assume X is dist with <u>empirical dist</u>: discrete uniform over sample pts
- Empirical risk: expected loss under empirical distribution

$$\hat{R}(h) = \frac{1}{n} \sum_{i=1}^{n} L(h(x_i), y_i)$$

why we minimize the <u>sum</u> of loss functions

Logistic Loss from max likelihood

what cost fn for probabilities?

prob pt. X_i in y_i ← actual prob

Make β duplicate X_i's, $y_i\beta$ in class $(1-y_i)\beta$ not

$$\mathcal{L}(h; X, y) = \prod_{i=1}^{n} h(x_i)^{y_i \beta} (1-h(x_i))^{(1-y_i)\beta}$$ ← almost binomial distribution

$$l(h) = \beta \sum_{i=1}^{n} y_i \ln h(x_i) + (1-y_i) \ln(1-h(x_i))$$

$$= -\beta \sum \text{logistic loss fn}$$

$\Rightarrow$ min all logistic losses

The bias-variance decomposition

bias: error due to inability of hypoth h to fit g
ex. fitting quadratic g with linear h

variance: error due to fitting random noise in data
ex. fit linear g with linear h, yet $h \neq g$

$X_i \sim D$, $\varepsilon_i \sim D'$, $y_i = g(x_i) + \varepsilon_i$ fit h

pt. $z \in \mathbb{R}^d$ $\quad y = g(z) + \varepsilon$

$\mathbb{E}[y] = g(z)$ $\quad \mathrm{Var}(y) = \mathrm{Var}(\varepsilon)$

$$R(h) = \mathbb{E}[L(h(z), y)] = \mathbb{E}[(h(z) - y)^2]$$

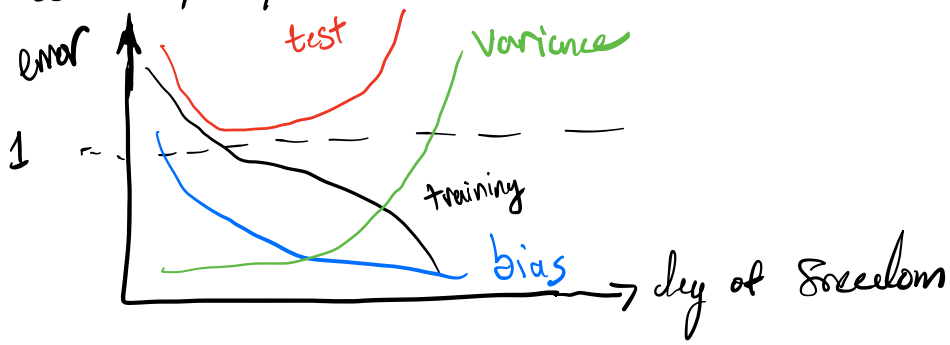
$$= \mathbb{E}[h(z)^2] + \mathbb{E}[y^2] - 2\,\mathbb{E}[y\,h(z)]$$

$$= \mathrm{Var}(h(z)) + \mathbb{E}[h(z)]^2 + \mathrm{Var}(y) + \mathbb{E}[y]^2 - 2\,\mathbb{E}[y]\,\mathbb{E}[h(z)]$$

$$= \underbrace{(\mathbb{E}[h(z)] - \mathbb{E}[y])^2}_{\text{bias}} + \underbrace{\mathrm{Var}(h(z))}_{\text{variance}} + \underbrace{\mathrm{Var}(\varepsilon)}_{\text{irreducible error}}$$

"bias-variance decomposition"

- underfitting → too much bias
- overfitting → too much variance
- Training error reflects bias but not variance
- many dists, $\delta \to 0$ as $n \to \infty$
- if n can be g, bias $\to 0$ " "
  ↳ if can't bias large
- good feature ↓ bias, bad rarely increases bias
  ↳ usually increases variance
- irreducible error can't be reduced
- noise in test affects $\text{Var}(\varepsilon)$
- noise in train bias / variance
- can't precisely measure bias/variance
  ↳ test w/ synthetic data



crossing point of derivative