Statistical Learning

$$P_{XY}(X = x, Y = y)$$

Goal: Predict Y given X

Find a function η that minimizes

$$\mathbb{E}_{XY}[(Y - \eta(X))^2]$$

a metric to evaluate a prediction function

P_{XY}

quadratic → regression

other losses: l1 logistic 0/1 loss

Thus far, we've been using η which is a:

- **Linear functions of X**
- **Degree p polynomials of X**
- **Linear "generalization" of X**

Statistical Learning

best / optimal predictor

$P_{XY}(X = x, Y = y)$

Goal: Predict Y given X

Find a function η that minimizes

$$\mathbb{E}_{XY}[(Y - \eta(X))^2] = \mathbb{E}_X\left[\mathbb{E}_{Y|X}[(Y - \eta(x))^2 | X = x]\right]$$

$$\eta(x) = \arg\min_c \mathbb{E}_{Y|X}[(Y - c)^2 | X = x] = \mathbb{E}_{Y|X}[Y | X = x]$$

Thm :

Under LS loss, optimal predictor: $\eta(x) = \mathbb{E}_{Y|X}[Y|X = x]$

Optimal Prediction

$$\mathbb{E}_{XY}[(Y-\eta(X))^2] = \mathbb{E}_X\Big[\mathbb{E}_{Y|X}[(Y-\eta(x))^2|X=x]\Big]$$

Under LS loss, optimal predictor: $\eta(x) = \mathbb{E}_{Y|X}[Y|X=x]$

pf: $0 = \frac{d}{d\eta(x)} \mathbb{E}_{Y|X}\left[(Y-\eta(x))^2 \,\middle|\, X=x\right]$

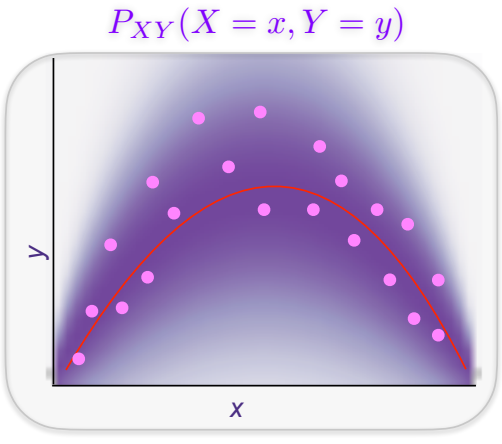
$= \mathbb{E}_{Y|X}\left[\frac{d}{d\eta(x)}(Y-\eta(x))^2 \,\middle|\, X=x\right]$

$= \mathbb{E}_{Y|X}\left[-2(Y-\eta(x)) \,\middle|\, X=x\right]$

$= -2\mathbb{E}_{Y|X}[Y|X=x] + 2\eta(x)$

$\Rightarrow \quad \eta(x) = \mathbb{E}_{Y|X}[Y|X=x]$

Statistical Learning



Ideally, we want to find:

$$\eta(x) = \mathbb{E}_{Y|X}[Y|X = x]$$

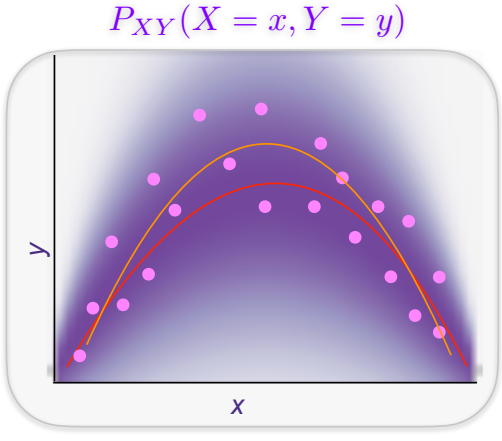
But we only have samples:

$(x_i, y_i) \overset{i.i.d.}{\sim} P_{XY}$ for $i = 1, \ldots, n$

$\{(x_i, y_i)\}_{i=1}^{n}$

$\Rightarrow$ only estimate $\eta(x)$

Statistical Learning



Ideally, we want to find:

$$\eta(x) = \mathbb{E}_{Y|X}[Y|X = x]$$

But we only have samples:

$$(x_i, y_i) \overset{i.i.d.}{\sim} P_{XY} \quad \text{for } i = 1, \ldots, n$$

and are restricted to a function class (e.g., linear) so we compute:

$$\widehat{f} = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2$$

We care about future predictions: $\mathbb{E}_{XY}[(Y - \widehat{f}(X))^2]$

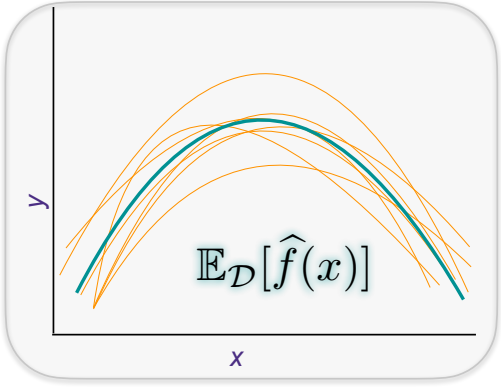
$\mathcal{F}$: predictor class

(1) linear
(2) quadratic
(3) degree-p poly
(4) generalized linear
(5) kernel
(6) neural network
.....

Q: is f a random or deterministic?

Statistical Learning

$P_{XY}(X = x, Y = y)$



Ideally, we want to find:

$$\eta(x) = \mathbb{E}_{Y|X}[Y|X = x]$$

But we only have samples:

$(x_i, y_i) \overset{i.i.d.}{\sim} P_{XY}$ for $i = 1, \ldots, n$

and are restricted to a function class (e.g., linear) so we compute:

$$\widehat{f} = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2$$

f is deterministic given D

Each draw $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ results in different $\widehat{f}$

Bias-Variance Tradeoff

$$D = \{(X_i, Y_i)\}_{i=1}^{n} \overset{\text{i.i.d.}}{\sim} P_{XY}$$

new $\tilde{x}$

$$\eta(x) = \mathbb{E}_{Y|X}[Y|X = x] \qquad \hat{f} = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2$$

$$\mathbb{E}_{Y|X}\mathbb{E}_D[(f - g(x))(g(x) - f_D(x))|X=x]$$
$$= \mathbb{E}_{Y|X}[(Y - g(x))(\mathbb{E}_D(g(x) - f_D(x)))|X=x]$$
(we know $\mathbb{E}_{Y|X}[Y] = g(x)$)
$$= 0$$

What we care $\mathbb{E}_{XY}[(Y - f(x))^2]$

$$\Rightarrow \mathbb{E}_{XY}[\mathbb{E}_D(Y - f_D(x))^2]$$
$$= \mathbb{E}_X[\mathbb{E}_{Y|X}\mathbb{E}_D(Y - f_D(x))^2 | X=x]$$
$$= \mathbb{E}_X[\mathbb{E}_{Y|X}\mathbb{E}_D[Y - g(x) + g(x) - f_D(x)]^2 | X=x]$$
$$= \mathbb{E}_X[\mathbb{E}_{Y|X}\mathbb{E}_D[(Y - g(x))^2 + 2(Y - g(x))(g(x) - f_D(x)) + (g(x) - f_D(x))^2 | X=x]$$
$$= \mathbb{E}_X[\mathbb{E}_{Y|X}(Y - g(x))^2] + \mathbb{E}_X[\mathbb{E}_{Y|X}\mathbb{E}_D(g(x) - f_D(x))^2]$$

Bias-Variance Tradeoff

$\eta(x) = \mathbb{E}_{Y|X}[Y|X = x]$ $\quad$ $\hat{f} = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2$

(1) $\mathbb{E}_{XY}\left[(Y - \eta(x))^2\right]$: irreducible error
independent of data
caused by stochastic

(2) $\mathbb{E}_{XY}\left[\mathbb{E}_D\left[(\eta(x) - f_D(x))^2\right]\right]$

$= \mathbb{E}_X\left[\mathbb{E}_D\left(\eta(x) - f_D(x)\right)^2\right]$:

learning error
depend on D
- $\mathcal{F}$ is too simple
- limited data

Bias-Variance Tradeoff

fix x

$D = R_{XY}^n$ $\quad D$: diff over $\{(x_i, y_i)\}_{i=1}^n$, x is independent of D

$$\mathbb{E}_D[f_D(x)] = \int_{\{(x_i,y_i)\}_{i=1}^n} f_{\{(x_i,y_i)\}_{i=1}^n}(x)\, d\mathcal{D}(\{(x_i,y_i)\}_{i=1}^n)$$

$$\eta(x) = \mathbb{E}_{Y|X}[Y|X=x] \qquad \hat{f} = \arg\min_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2$$

learning error

fix x:

$$\mathbb{E}_D\left[(\eta_0(x) - f_0(x))^2\right]$$

$$= \mathbb{E}_D\left[(\eta_0(x) - \mathbb{E}_D[f_D(x)] + \mathbb{E}_D[f_D(x)] - f_D(x))^2\right]$$

$$= \mathbb{E}_D\Big[(\eta_0(x) - \mathbb{E}_D[f_D(x)])^2 + 2\underbrace{(\eta_0(x) - \mathbb{E}_D[f_D(x)])(\mathbb{E}_D[f_D(x)] - f_D(x))}_{=0} + (\mathbb{E}_D[f_D(x)] - f_D(x))^2\Big]$$

$$= (\eta_0(x) - \mathbb{E}_D[f_D(x)])^2 + \mathbb{E}_D\left[(\mathbb{E}_D[f_D(x)] - f_D(x))^2\right]$$

bias squared

variance

$f_D(x)$: a random variable

independent of η_0

$P_{XY} \Rightarrow \eta(x) = \mathbb{E}[Y \mid X = x]$

$D = P_{XY}^{n}$

$\{(x_i, y_i)\}_{i=1}^{n} \sim D$

$\Rightarrow \quad f_D$

$\mathbb{E}_D[f_D]$

Bias-Variance Tradeoff

more complex model
small bias
$\hat{f} \leftarrow \text{argmin}_{f \in \mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}(f(x_i)-y_i)^2$
$\exists f \approx y$

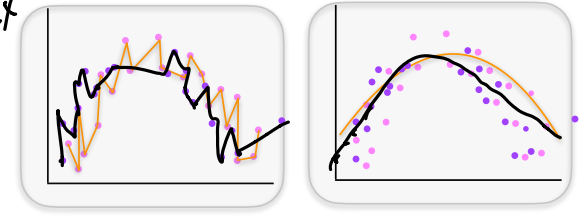
$$\mathbb{E}_{Y|X}[\mathbb{E}_{\mathcal{D}}[(Y-\hat{f}_{\mathcal{D}}(x))^2]|X=x] = \mathbb{E}_{Y|X}[(Y-\eta(x))^2|X=x]$$

irreducible error

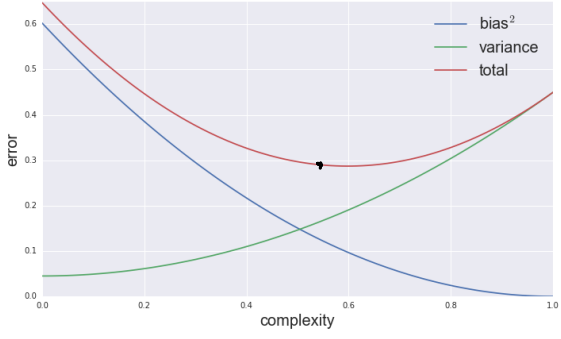
$$+(\eta(x)-\mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)])^2 + \mathbb{E}_{\mathcal{D}}[(\mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)]-\hat{f}_{\mathcal{D}}(x))^2]$$

bias squared **variance**

less complex
→ smaller variance



If we re-drew our data, what the LS training error estimator look like for generalized linear functions in small p/large p dimensions?



How many points do I use for training/testing?

- **Very hard question to answer!**
  - **Too few training points, learned model is bad**
  - **Too few test points, you never know if you reached a good solution**
- **Bounds, such as Hoeffding's inequality can help:**

$$P(|\ \widehat{\theta} - \theta^* \ | \geq \epsilon) \ \leq \ 2e^{-2N\epsilon^2}$$

- **More on this later the quarter, but still hard to answer**
- **Typically:**
  - **If you have a reasonable amount of data 90/10 splits are common**
  - **If you have little data, then you need to get fancy (e.g., bootstrapping)**