Linear Regression

Machine Learning – CSE546
Kevin Jamieson
University of Washington

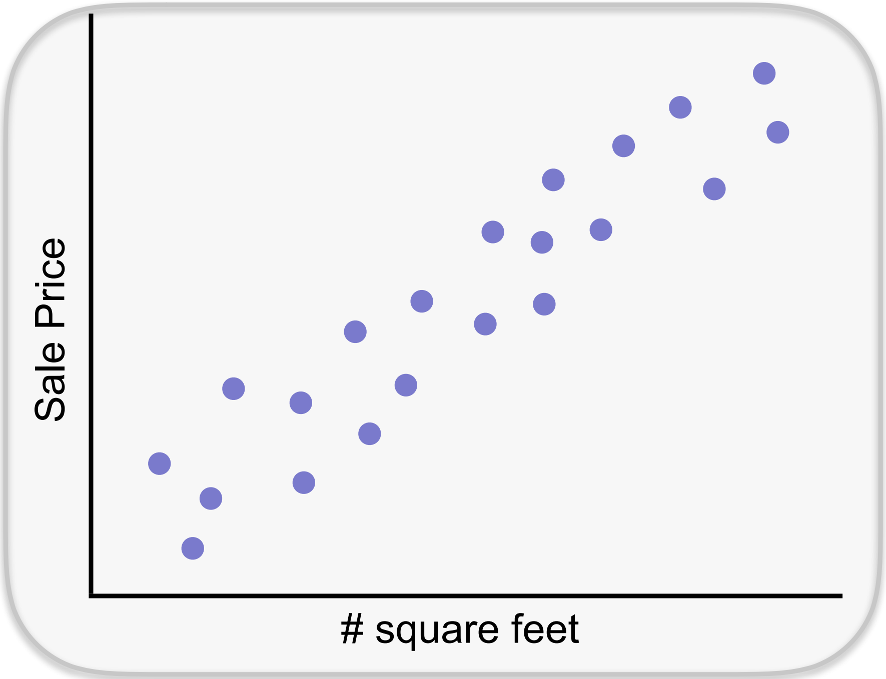
Oct 2, 2018

The regression problem

Given past sales data on zillow.com, predict:

y = **House sale price** *from*

x = **{# sq. ft., zip code, date of sale, etc.}**



Training Data:

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

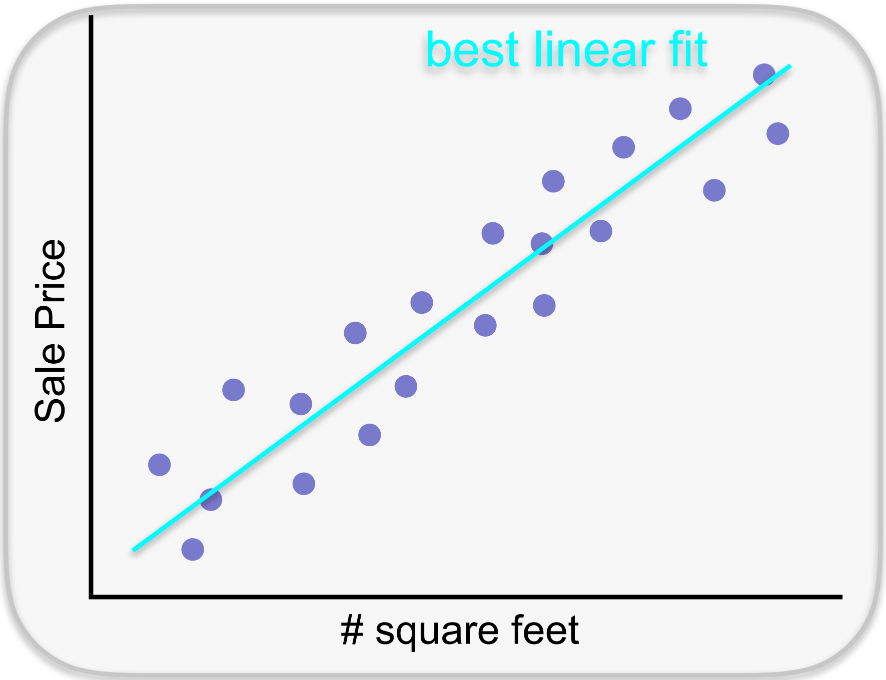
$$x_i \in \mathbb{R}^d$$
$$y_i \in \mathbb{R}$$

The regression problem

Given past sales data on zillow.com, predict:

y = **House sale price** *from*

x = **{# sq. ft., zip code, date of sale, etc.}**



Training Data: $\{(x_i, y_i)\}_{i=1}^n$ $\quad x_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}$

Hypothesis: linear

$$y_i \approx x_i^T w$$

Loss: least squares

$$\min_w \sum_{i=1}^n \left(y_i - x_i^T w\right)^2$$

The regression problem in matrix notation

$$
\begin{aligned}
\widehat{w}_{LS} &= \arg\min_w \sum_{i=1}^n \left(y_i - x_i^T w\right)^2 \\
&= \arg\min_w (\mathbf{y} - \mathbf{X}w)^T (\mathbf{y} - \mathbf{X}w)
\end{aligned}
$$

$$
\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix}
$$

The regression problem in matrix notation

$$\widehat{w}_{LS} = \arg\min_w ||\mathbf{y} - \mathbf{X}w||_2^2$$
$$= \arg\min_w (\mathbf{y} - \mathbf{X}w)^T(\mathbf{y} - \mathbf{X}w)$$

$$\nabla_w(\cdot) = -2X^T(y - Xw) = 0$$

$$\boxed{X^T y = X^T X w}$$

If $(X^TX)^{-1}$ exists then $\hat{w} = (X^TX)^{-1}X^Ty$

The regression problem in matrix notation

$$\widehat{w}_{LS} = \arg\min_{w} ||\mathbf{y} - \mathbf{X}w||_2^2$$
$$= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

What about an offset?

$$\widehat{w}_{LS}, \widehat{b}_{LS} = \arg\min_{w,b} \sum_{i=1}^{n} \left(y_i - (x_i^T w + b)\right)^2$$
$$= \arg\min_{w,b} ||\mathbf{y} - (\mathbf{X}w + \mathbf{1}b)||_2^2$$

Dealing with an offset

$$\mathbf{1} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$

$$\widehat{w}_{LS}, \widehat{b}_{LS} = \arg\min_{w,b} ||\mathbf{y} - (\mathbf{X}w + \mathbf{1}b)||_2^2$$

$$\nabla_w = 2(X^T)(y - (Xw + \mathbf{1}b)) = 0$$

$$X^T y = X^T X w + X^T \mathbf{1} b$$

$$\nabla_b = 2\,\mathbf{1}^T (y - (Xw + \mathbf{1}b)) = 0$$

$$\mathbf{1}^T \mathbf{1} = n$$

$$\mathbf{1}^T y = \mathbf{1}^T X w + n b$$

$$\mathbf{1}^T y = \sum_{i=1}^n y_i$$

$$b = \frac{1}{n} \sum y_i - \frac{1}{n} \mathbf{1}^T X w$$

Dealing with an offset

$$\widehat{w}_{LS}, \widehat{b}_{LS} = \arg\min_{w,b} ||\mathbf{y} - (\mathbf{X}w + \mathbf{1}b)||_2^2$$

$$\mathbf{X}^T\mathbf{X}\widehat{w}_{LS} + \widehat{b}_{LS}\mathbf{X}^T\mathbf{1} = \mathbf{X}^T\mathbf{y}$$

$$\mathbf{1}^T\mathbf{X}\widehat{w}_{LS} + \widehat{b}_{LS}\mathbf{1}^T\mathbf{1} = \mathbf{1}^T\mathbf{y}$$

$(\mathbf{1}^T X w)^T = w^T \underbrace{X^T \mathbf{1}}_{=0}$ $= 0$

$\boxed{\mu = \frac{1}{n}\sum_{i=1}^n x_i}$

$0 = \sum_{i=1}^n (x_i - \mu) = \left(\sum x_i\right) - n\mu = 0$

If $\underbrace{\mathbf{X}^T\mathbf{1}}_{=\sum_{i=1}^n x_i} = 0$ (i.e., if each feature is mean-zero) then

$$\widehat{w}_{LS} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

$\left.\begin{array}{l} \tilde{x}_i = x_i - \mu \\ \tilde{X} = \begin{bmatrix} \tilde{x}_1^T \\ \vdots \\ \tilde{x}_n^T \end{bmatrix} \end{array}\right] \to \widehat{w}$

$$\widehat{b}_{LS} = \frac{1}{n}\sum_{i=1}^n y_i$$

Given a new x, predict $\widehat{w}^T(x-\mu) + \widehat{b}$

The regression problem in matrix notation

$$\widehat{w}_{LS} = \arg\min_w ||\mathbf{y} - \mathbf{X}w||_2^2$$
$$= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

But why least squares?

Consider $\quad y_i = x_i^T w + \epsilon_i \quad$ where $\quad \epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$

$$P(y|x, w, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y - x^T w)^2}{2\sigma^2}\right)$$

Maximizing log-likelihood

Maximize:

$$\log P(\mathcal{D}|w,\sigma) = \log(\frac{1}{\sqrt{2\pi}\sigma})^n \prod_{i=1}^{n} e^{-\frac{(y_i - x_i^T w)^2}{2\sigma^2}}$$

MLE is LS under linear model

$$\widehat{w}_{LS} = \arg\min_w \sum_{i=1}^n \left(y_i - x_i^T w\right)^2$$

$$\widehat{w}_{MLE} = \arg\max_w P(\mathcal{D}|w,\sigma)$$

$$\text{if} \quad y_i = x_i^T w + \epsilon_i \quad \text{and} \quad \epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0,\sigma^2)$$

$$\boxed{\widehat{w}_{LS} = \widehat{w}_{MLE} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}}$$

Analysis of error

$$\underline{\mathbf{Y} = \mathbf{X}w + \epsilon}$$

if $\quad y_i = x_i^T w + \epsilon_i \quad$ and $\quad \boxed{\epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)}$

$$\widehat{w}_{MLE} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

$$= \underbrace{(X^TX)^{-1}X^T(Xw}_{} + \varepsilon)$$

$$= w + (X^TX)^{-1}X^T\varepsilon$$

$$\boxed{\mathbb{E}[\varepsilon\varepsilon^T] = \sigma^2 I}$$

$$\mathbb{E}[\hat{w}] = w$$

$$\mathbb{E}[(\hat{w}-w)(\hat{w}-w)^T] = \mathbb{E}[(X^TX)^{-1}X^T\varepsilon\varepsilon^T X(X^TX)^{-1}]$$

$$= (X^TX)^{-1}\underbrace{X^T\mathbb{E}[\varepsilon\varepsilon^T]X}_{}(X^TX)^{-1}$$

$$= \sigma^2(X^TX)^{-1}$$

$$\hat{w} \sim \mathcal{N}(w, \sigma^2(X^TX)^{-1})$$

Analysis of error $\mathbf{Y} = \mathbf{X}w + \epsilon$

if $\quad y_i = x_i^T w + \epsilon_i \quad$ and $\quad \epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$

$$\begin{aligned}
\widehat{w}_{MLE} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \\
&= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}w + \epsilon) \\
&= w + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\epsilon
\end{aligned}$$

$$\mathrm{Cov}(\widehat{w}_{MLE}) = \mathbb{E}[(\widehat{w} - \mathbb{E}[\widehat{w}])(\widehat{w} - \mathbb{E}[\widehat{w}])^T] = (\mathbf{X}^T\mathbf{X})^{-1}$$

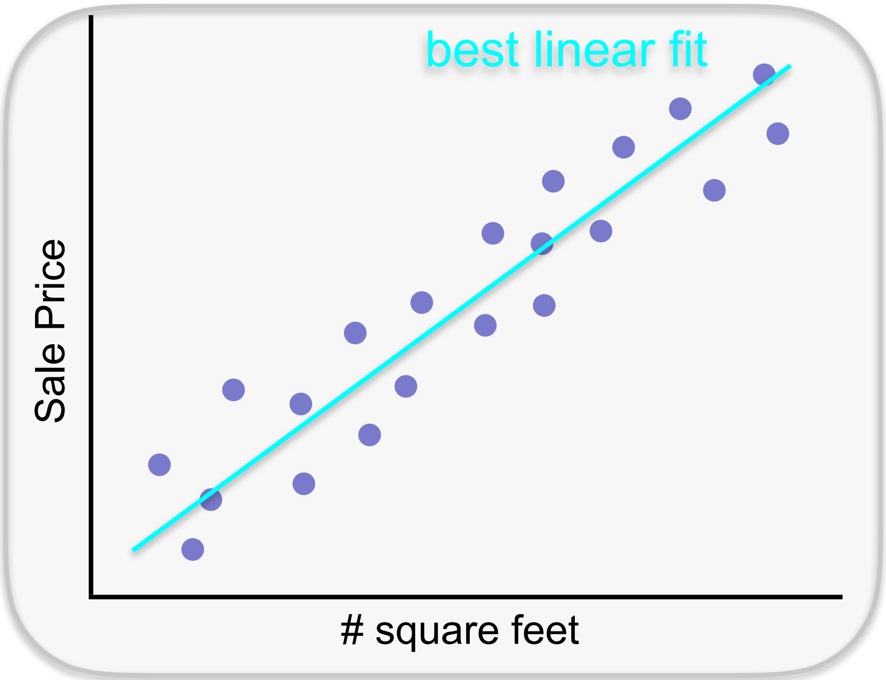
$$\widehat{w}_{MLE} \sim \mathcal{N}(w, (\mathbf{X}^T\mathbf{X})^{-1})$$

The regression problem

Given past sales data on zillow.com, predict:

y = **House sale price** *from*

x = **{# sq. ft., zip code, date of sale, etc.}**



Training Data:

$$\{(x_i, y_i)\}_{i=1}^n \qquad x_i \in \mathbb{R}^d \quad y_i \in \mathbb{R}$$

Hypothesis: linear

$$y_i \approx x_i^T w$$

Loss: least squares

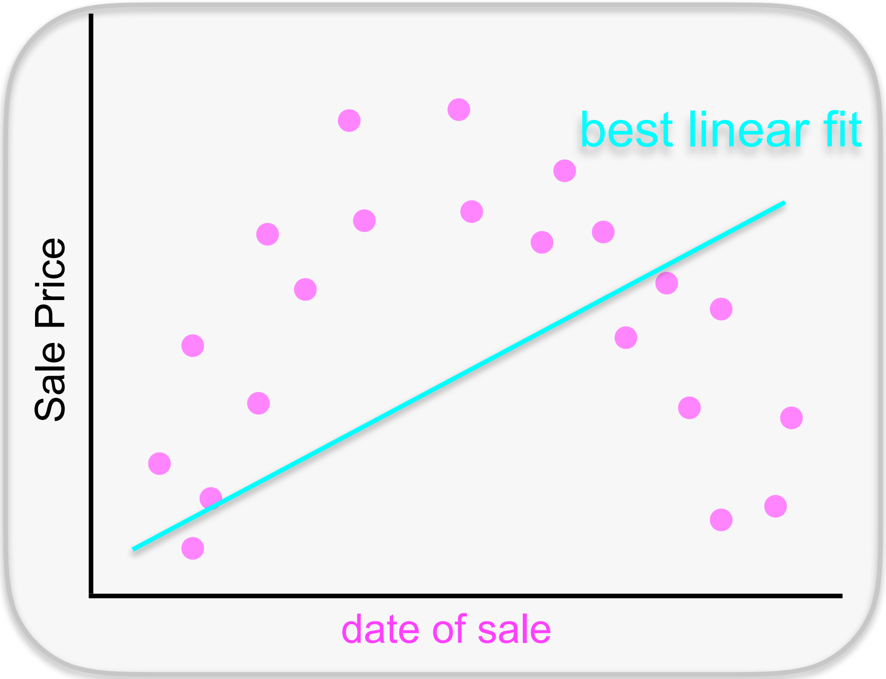
$$\min_w \sum_{i=1}^n \left(y_i - x_i^T w\right)^2$$

The regression problem

Given past sales data on zillow.com, predict:

y = **House sale price** *from*

x = **{# sq. ft., zip code, date of sale, etc.}**



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

$x_i \in \mathbb{R}^d$
$y_i \in \mathbb{R}$

Hypothesis: linear

$$y_i \approx x_i^T w$$

Loss: least squares

$$\min_w \sum_{i=1}^n \left(y_i - x_i^T w\right)^2$$

The regression problem

Training Data:

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

$x_i \in \mathbb{R}^d$
$y_i \in \mathbb{R}$

Transformed data:

Hypothesis: linear

$$y_i \approx x_i^T w$$

Loss: least squares

$$\min_w \sum_{i=1}^n \left(y_i - x_i^T w\right)^2$$

The regression problem

Training Data:

$$\{(x_i, y_i)\}_{i=1}^n \qquad \begin{matrix} x_i \in \mathbb{R}^d \\ y_i \in \mathbb{R} \end{matrix}$$

Hypothesis: linear

$$y_i \approx x_i^T w$$

Loss: least squares

$$\min_w \sum_{i=1}^n \left(y_i - x_i^T w\right)^2$$

Transformed data:

$h : \mathbb{R}^d \to \mathbb{R}^p$ maps original features to a rich, possibly high-dimensional space

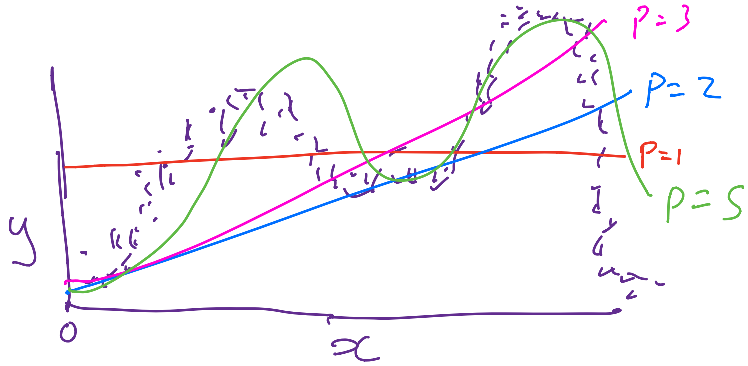
in d=1: $h(x) = \begin{bmatrix} h_1(x) \\ h_2(x) \\ \vdots \\ h_p(x) \end{bmatrix} = \begin{bmatrix} x \\ x^2 \\ \vdots \\ x^p \end{bmatrix}$

for d>1, generate $\{u_j\}_{j=1}^p \subset \mathbb{R}^d$

$$h_j(x) = \frac{1}{1 + \exp(u_j^T x)}$$

$$h_j(x) = (u_j^T x)^2$$

$$h_j(x) = \cos(u_j^T x)$$



$$\hat{f}(x) = \sum_{k=0}^{p-1} \underbrace{f^{(k)}(0)} \frac{1}{k!} x^k$$

$$h(x) = \begin{bmatrix} h_0(x) \\ \vdots \\ h_{p-1}(x) \end{bmatrix}$$

Given x, predict $\hat{w}^T h(x)$

$$h_k(x) = x^k \frac{1}{k!}$$

$$x_i \longmapsto h(x_i)$$

$$\breve{X} = \begin{bmatrix} h(x_1)^T \\ \vdots \\ h(x_n)^T \end{bmatrix}$$

$$\hat{w} = (\breve{X}^T \breve{X})^{-1} \breve{X}^T y$$

The regression problem

Training Data: $x_i \in \mathbb{R}^d$

$\{(x_i, y_i)\}_{i=1}^n$ $\quad y_i \in \mathbb{R}$

~~Hypothesis: linear~~

$$\cancel{y_i \approx x_i^T w}$$

~~Loss: least squares~~

$$\cancel{\min_w \sum_{i=1}^n \left(y_i - x_i^T w\right)^2}$$

Transformed data:

$$h(x) = \begin{bmatrix} h_1(x) \\ h_2(x) \\ \vdots \\ h_p(x) \end{bmatrix}$$

Hypothesis: linear

$$y_i \approx h(x_i)^T w \qquad w \in \mathbb{R}^p$$

Loss: least squares

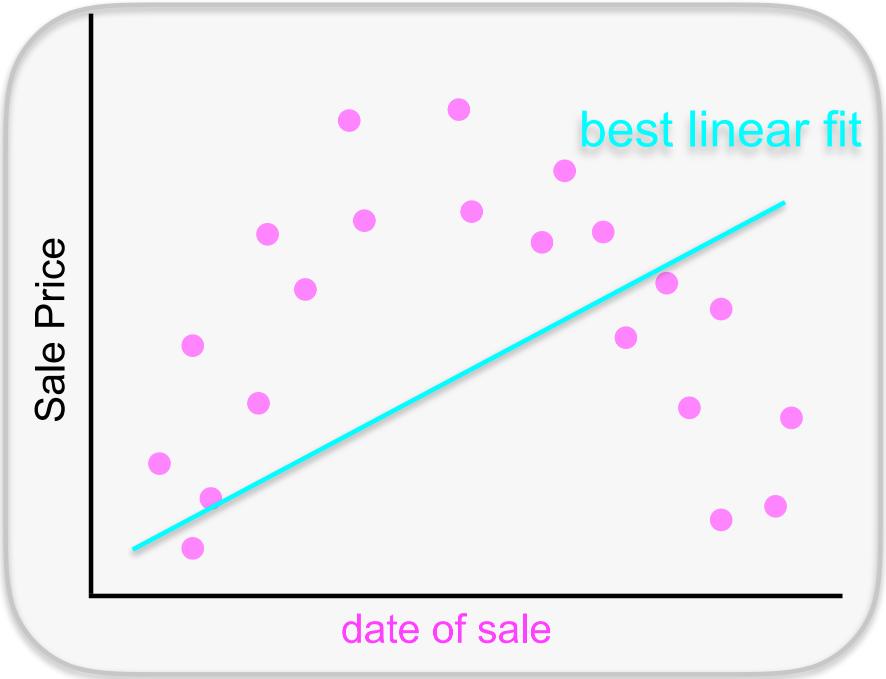
$$\min_w \sum_{i=1}^n \left(y_i - h(x_i)^T w\right)^2$$

The regression problem

Training Data:

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

$$x_i \in \mathbb{R}^d$$
$$y_i \in \mathbb{R}$$



Transformed data:

$$h(x) = \begin{bmatrix} h_1(x) \\ h_2(x) \\ \vdots \\ h_p(x) \end{bmatrix}$$

Hypothesis: linear

$$y_i \approx h(x_i)^T w \quad w \in \mathbb{R}^p$$

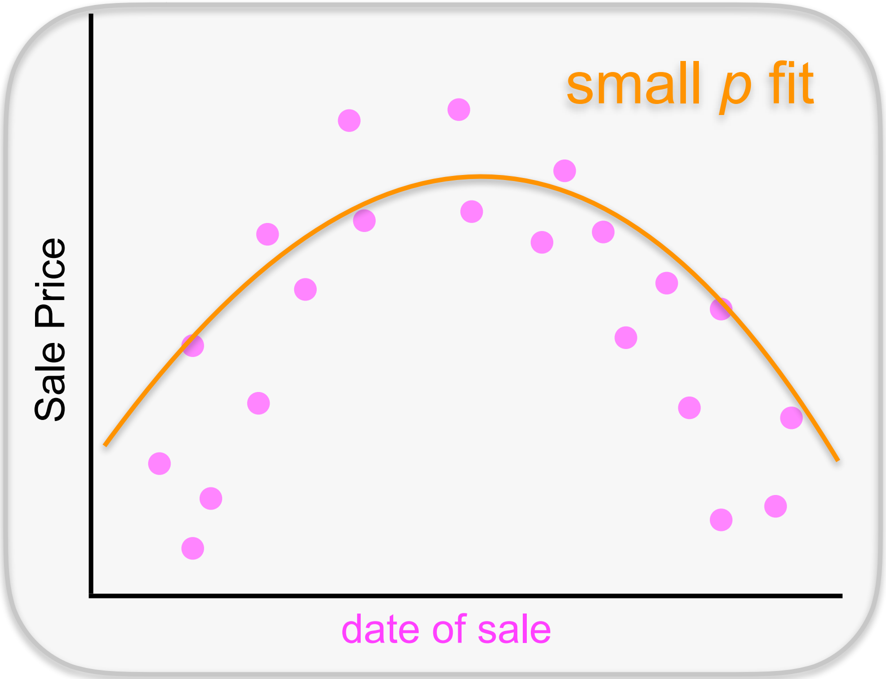
Loss: least squares

$$\min_w \sum_{i=1}^{n} \left(y_i - h(x_i)^T w\right)^2$$

The regression problem

Training Data:

$$\{(x_i, y_i)\}_{i=1}^n \qquad x_i \in \mathbb{R}^d \quad y_i \in \mathbb{R}$$



Transformed data:

$$h(x) = \begin{bmatrix} h_1(x) \\ h_2(x) \\ \vdots \\ h_p(x) \end{bmatrix}$$

Hypothesis: linear

$$y_i \approx h(x_i)^T w \quad w \in \mathbb{R}^p$$

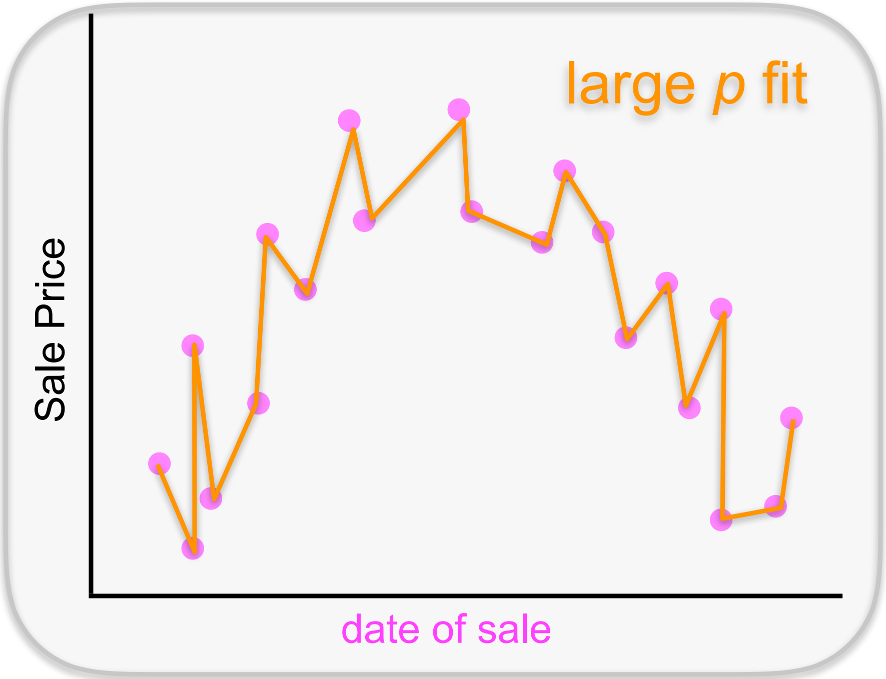
Loss: least squares

$$\min_w \sum_{i=1}^n \left(y_i - h(x_i)^T w\right)^2$$

The regression problem

Training Data:

$$\{(x_i, y_i)\}_{i=1}^n \qquad x_i \in \mathbb{R}^d \quad y_i \in \mathbb{R}$$



Transformed data:

$$h(x) = \begin{bmatrix} h_1(x) \\ h_2(x) \\ \vdots \\ h_p(x) \end{bmatrix}$$

Hypothesis: linear

$$y_i \approx h(x_i)^T w \quad w \in \mathbb{R}^p$$

Loss: least squares

$$\min_w \sum_{i=1}^n \left(y_i - h(x_i)^T w\right)^2$$

What's going on here?

Bias-Variance Tradeoff

Machine Learning – CSE546
Kevin Jamieson
University of Washington

Oct 5, 2018

Statistical Learning

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

Goal: Predict Y given X

Find function η that minimizes

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

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

$$\frac{d}{dc}\mathbb{E}_{Y|X}\left[(Y - c)^2 \mid X = x\right] = \mathbb{E}\left[2(Y - c) \mid X = x\right] = 0$$

$$= 2\,\mathbb{E}[Y \mid X = x] - 2c = 0$$

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

Statistical Learning

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

Goal: Predict Y given X

Find function η that minimizes

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

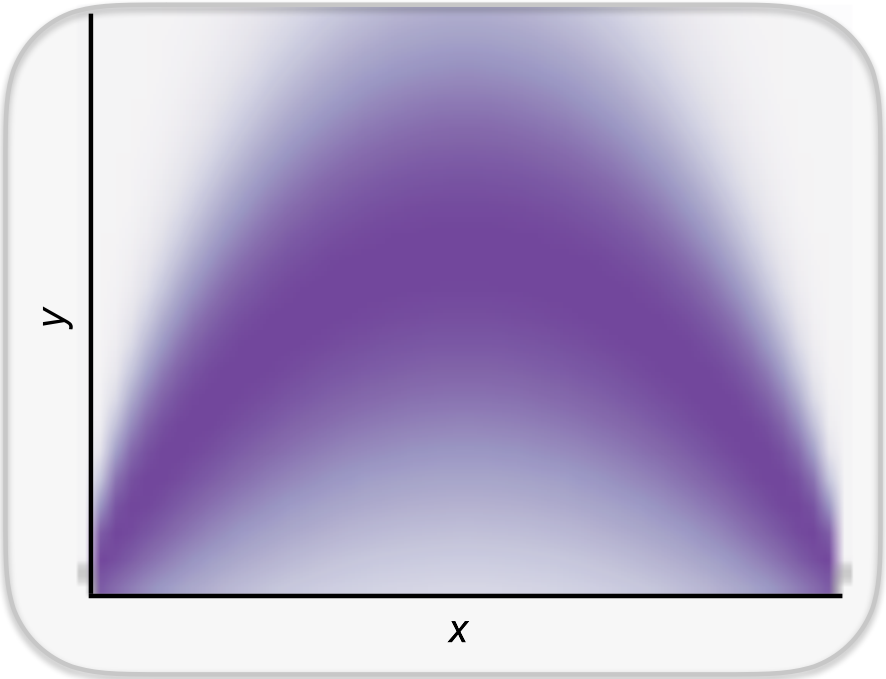
$$\eta(x) = \arg\min_c \mathbb{E}_{Y|X}[(Y - c)^2 | X = x] = \mathbb{E}_{Y|X}[Y | X = x]$$

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

Statistical Learning

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

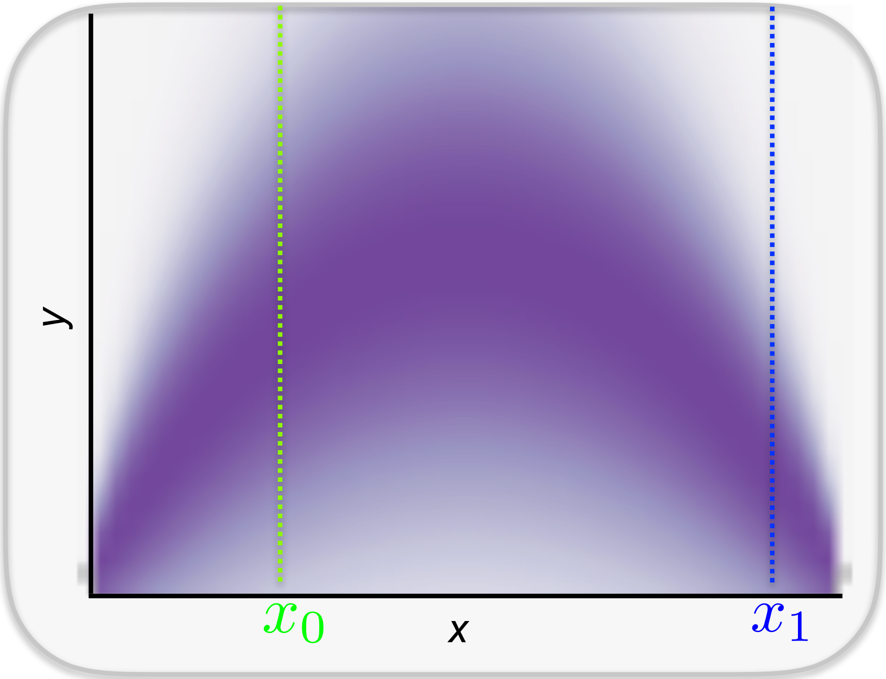
$$P_{XY}(X = x, Y = y)$$



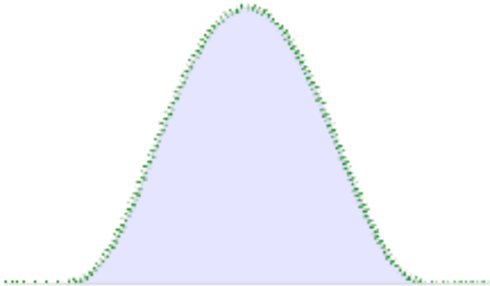
Statistical Learning

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

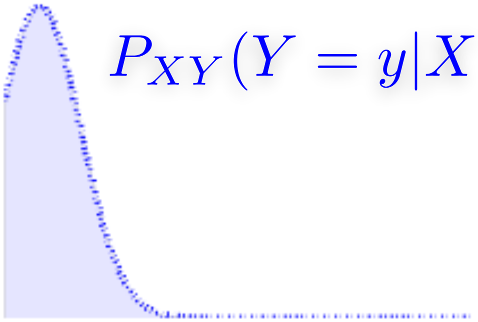
$P_{XY}(X=x, Y=y)$



$P_{XY}(Y=y|X=x_0)$



$P_{XY}(Y=y|X=x_1)$

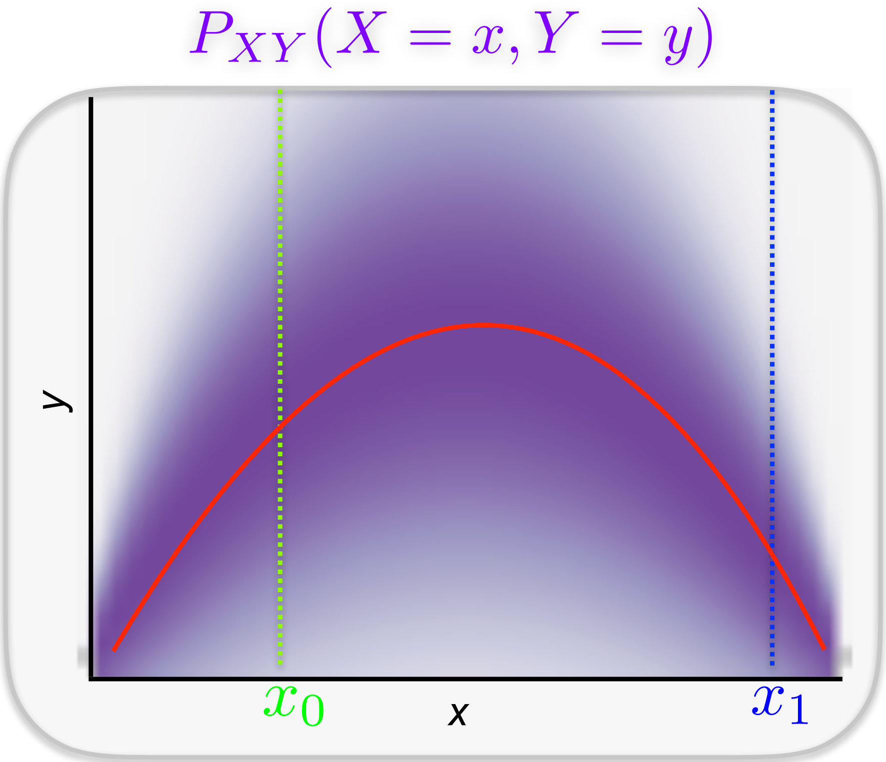


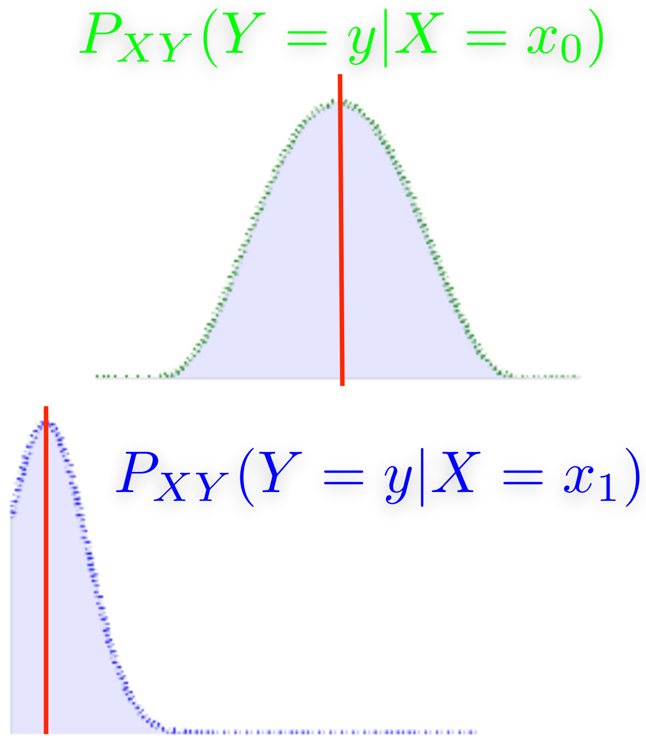
Statistical Learning

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

Ideally, we want to find:

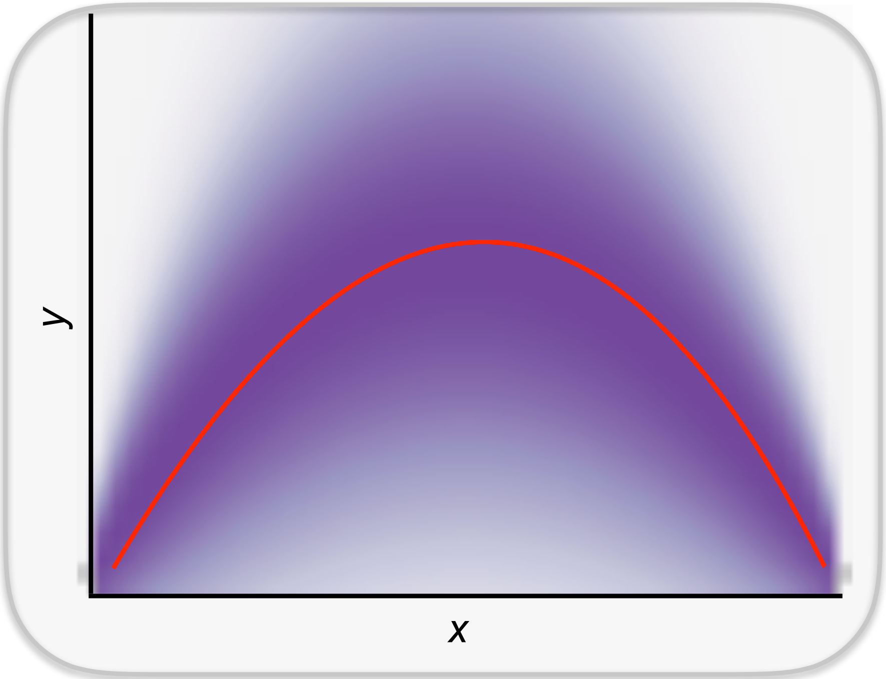
$$\eta(x) = \mathbb{E}_{Y|X}[Y|X = x]$$





Statistical Learning

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

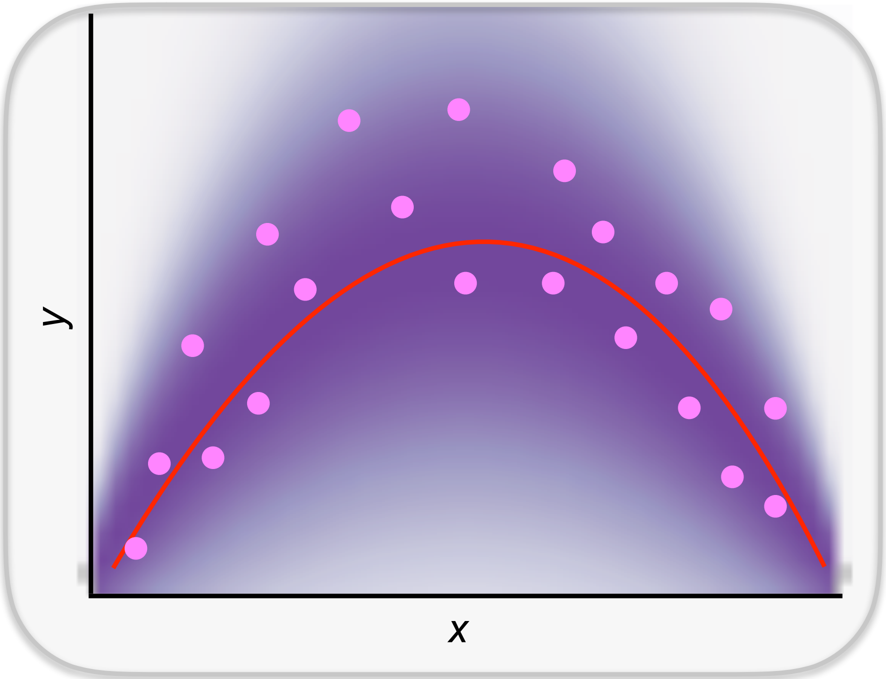


Ideally, we want to find:

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

Statistical Learning

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



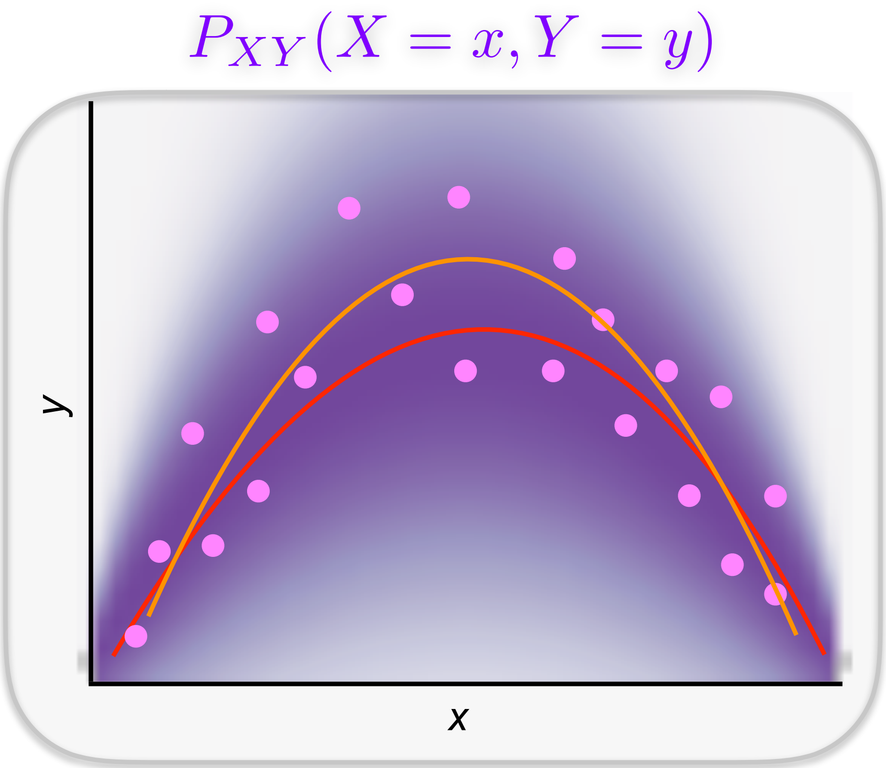
Ideally, we want to find:

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

But we only have samples:

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

Statistical Learning



Ideally, we want to find:

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

But we only have samples:

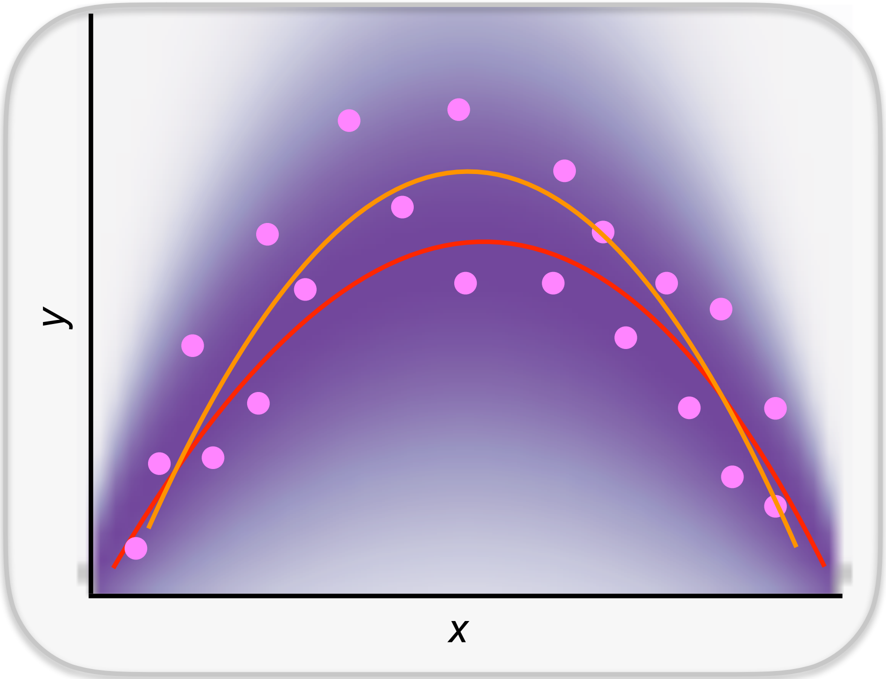
$(x_i, y_i) \overset{i.i.d.}{\sim} P_{XY}$ for $i = 1, \ldots, n$

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

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

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 $i = 1, \ldots, n$

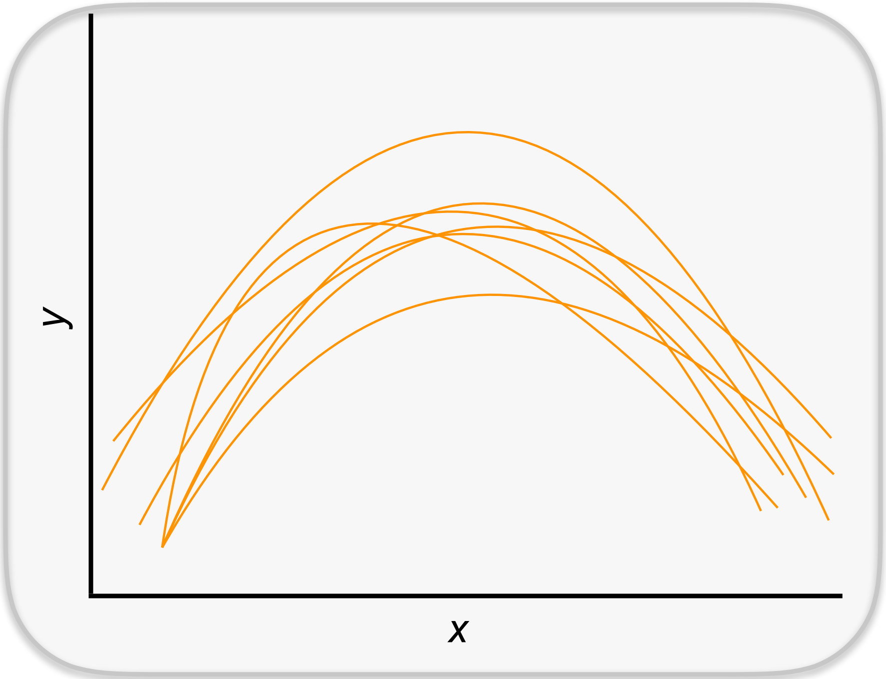
and are restricted to a function class (e.g., linear) so we compute:

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

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

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 $i = 1, \dots, n$

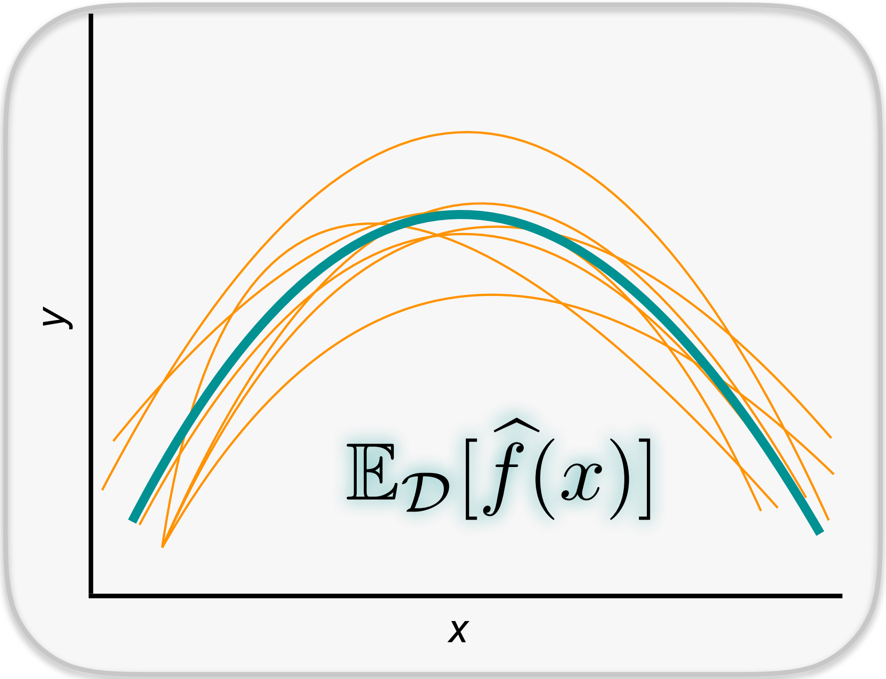
and are restricted to a function class (e.g., linear) so we compute:

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

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

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 $i = 1, \dots, n$

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

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

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

Bias-Variance Tradeoff

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

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

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

$$= \mathbb{E}_{Y|X}\Big[(Y-\eta(x))^2 \Big| X=x\Big] + \cancelto{0}{2\,\mathbb{E}_D\Big[\mathbb{E}_{Y|X}\big[(Y-\eta(x))(\eta(x)-\hat{f}_D(x))\big]\Big]} + \mathbb{E}_D\Big[(\eta(x)-\hat{f}_D(x))^2\Big]$$

$$\mathbb{E}_{Y|X}\big[(Y-\eta(x)) \,|\, X=x\big] = \mathbb{E}_{Y|X}\big[Y \,|\, X=x\big] - \eta(x) = 0$$

Bias-Variance Tradeoff

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

$$\begin{aligned}
\mathbb{E}_{Y|X}[\mathbb{E}_{\mathcal{D}}[(Y-\widehat{f}_{\mathcal{D}}(x))^2]\big|X=x] &= \mathbb{E}_{Y|X}[\mathbb{E}_{\mathcal{D}}[(Y-\eta(x)+\eta(x)-\widehat{f}_{\mathcal{D}}(x))^2]\big|X=x] \\
&= \mathbb{E}_{Y|X}\Big[\mathbb{E}_{\mathcal{D}}[(Y-\eta(x))^2 + 2(Y-\eta(x))(\eta(x)-\widehat{f}_{\mathcal{D}}(x)) \\
&\quad + (\eta(x)-\widehat{f}_{\mathcal{D}}(x))^2]\big|X=x\Big] \\
&= \underbrace{\mathbb{E}_{Y|X}[(Y-\eta(x))^2\big|X=x]}_{\text{irreducible error}} + \underbrace{\mathbb{E}_{\mathcal{D}}[(\eta(x)-\widehat{f}_{\mathcal{D}}(x))^2]}_{\text{learning error}}
\end{aligned}$$

irreducible error
Caused by stochastic label noise

learning error
Caused by either using too "simple" of a model or not enough data to learn the model accurately

Bias-Variance Tradeoff

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

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

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

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

Bias-Variance Tradeoff

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

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

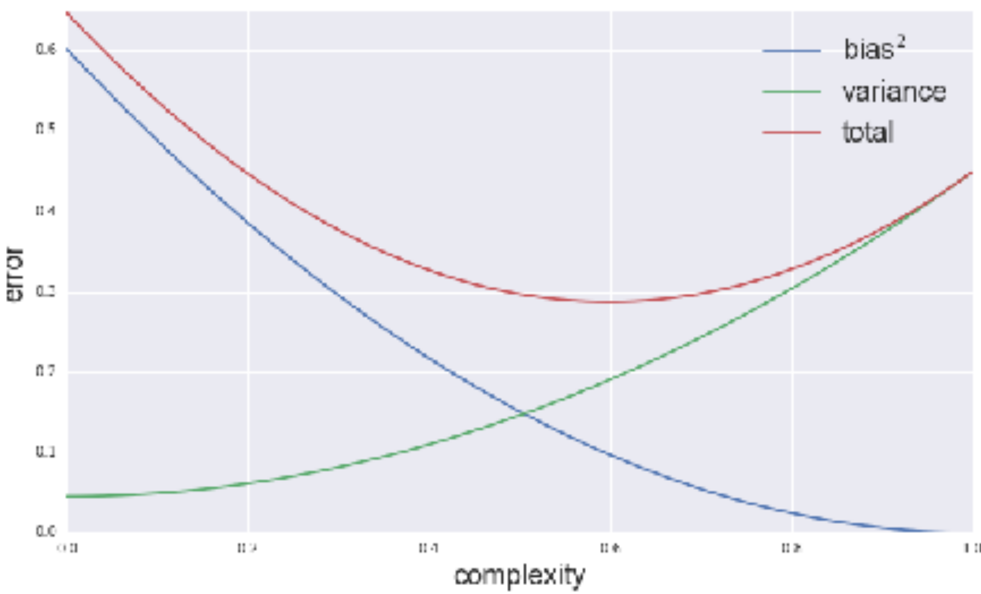
$$= \mathbb{E}_{\mathcal{D}}[(\eta(x) - \mathbb{E}_{\mathcal{D}}[\widehat{f}_{\mathcal{D}}(x)])^2 + 2(\eta(x) - \mathbb{E}_{\mathcal{D}}[\widehat{f}_{\mathcal{D}}(x)])(\mathbb{E}_{\mathcal{D}}[\widehat{f}_{\mathcal{D}}(x)] - \widehat{f}_{\mathcal{D}}(x))$$
$$+ (\mathbb{E}_{\mathcal{D}}[\widehat{f}_{\mathcal{D}}(x)]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

$$= \underbrace{(\eta(x) - \mathbb{E}_{\mathcal{D}}[\widehat{f}_{\mathcal{D}}(x)])^2}_{\textbf{biased squared}} + \underbrace{\mathbb{E}_{\mathcal{D}}[(\mathbb{E}_{\mathcal{D}}[\widehat{f}_{\mathcal{D}}(x)] - \widehat{f}_{\mathcal{D}}(x))^2]}_{\textbf{variance}}$$

Bias-Variance Tradeoff

$$\mathbb{E}_{Y|X}[\mathbb{E}_{\mathcal{D}}[(Y - \widehat{f}_{\mathcal{D}}(x))^2]\big|X = x] = \underbrace{\mathbb{E}_{Y|X}[(Y - \eta(x))^2\big|X = x]}_{\textbf{irreducible error}}$$

$$+\underbrace{(\eta(x) - \mathbb{E}_{\mathcal{D}}[\widehat{f}_{\mathcal{D}}(x)])^2}_{\textbf{biased squared}} + \underbrace{\mathbb{E}_{\mathcal{D}}[(\mathbb{E}_{\mathcal{D}}[\widehat{f}_{\mathcal{D}}(x)] - \widehat{f}_{\mathcal{D}}(x))^2]}_{\textbf{variance}}$$



Example: Linear LS $\mathbf{Y} = \mathbf{X}w + \epsilon$

if $\quad y_i = x_i^T w + \epsilon_i \quad$ and $\quad \epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$

$$\widehat{w}_{MLE} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} = w + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\epsilon$$

$$\eta(x) = \mathbb{E}_{Y|X}[Y|X=x] = \mathbb{E}[x^T w + \varepsilon \mid X = x] = x^T w$$

$$\widehat{f}_{\mathcal{D}}(x) = \hat{w}^T x$$

$$\mathbb{E}_D[\hat{f}_D(x)] = \mathbb{E}_{D|X}[\mathbb{E}_{Y|X}[\hat{w}^T x \mid X = x]] = \mathbb{E}_{D|X}[w^T x] = w^T x$$

Example: Linear LS $\mathbf{Y} = \mathbf{X}w + \epsilon$

if $\quad y_i = x_i^T w + \epsilon_i \quad$ and $\quad \epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$

$$\widehat{w}_{MLE} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} = w + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\epsilon$$

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

$$\widehat{f}_{\mathcal{D}}(x) = \widehat{w}^T x = w^T x + \epsilon^T \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} x$$

$$\mathbb{E}_{XY}[(\overbrace{Y - \eta(x)}^{\varepsilon})^2 | X = x] = \sigma^2 \qquad (\eta(x) - \mathbb{E}_{\mathcal{D}}[\widehat{f}_{\mathcal{D}}(x)])^2 = 0$$

irreducible error **biased squared**

Example: Linear LS $\mathbf{Y} = \mathbf{X}w + \epsilon$

if $\quad y_i = x_i^T w + \epsilon_i \quad$ and $\quad \epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$

$$\widehat{w}_{MLE} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} = w + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\epsilon$$

$$\widehat{f}_{\mathcal{D}}(x) = \underline{\widehat{w}^T x = w^T x + \epsilon^T \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}x}$$

$$\underbrace{\mathbb{E}_{\mathcal{D}}[(\mathbb{E}_{\mathcal{D}}[\widehat{f}_{\mathcal{D}}(x)] - \widehat{f}_{\mathcal{D}}(x))^2]}_{\textbf{variance}} = \mathbb{E}_D\left[x^T (X^T X)^{-1} X^T \underline{\varepsilon \varepsilon^T} X (X^T X)^{-1} x\right]$$

$$= \mathbb{E}_{X_D}\left[x^T (X^T X)^{-1} x\right] \sigma^2$$

Example: Linear LS $\mathbf{Y} = \mathbf{X}w + \epsilon$

if $\quad y_i = x_i^T w + \epsilon_i \quad$ and $\quad \epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$

$$\widehat{w}_{MLE} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} = w + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\epsilon$$

$$\widehat{f}_{\mathcal{D}}(x) = \widehat{w}^T x = w^T x + \epsilon^T \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}x$$

$$\underbrace{\mathbb{E}_{\mathcal{D}}[(\mathbb{E}_{\mathcal{D}}[\widehat{f}_{\mathcal{D}}(x)] - \widehat{f}_{\mathcal{D}}(x))^2]}_{\textbf{variance}} = \mathbb{E}_{\mathcal{D}}[x^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\epsilon\epsilon^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}x]$$

$$= \sigma^2 x^T(\mathbf{X}^T\mathbf{X})^{-1}x$$

$$= \sigma^2 \mathrm{Trace}((\mathbf{X}^T\mathbf{X})^{-1}xx^T)$$

$$\mathbf{X}^T\mathbf{X} = \sum_{i=1}^n x_i x_i^T \overset{n \text{ large}}{\longrightarrow} n\Sigma \qquad \Sigma = \mathbb{E}[XX^T], \quad X \sim P_X$$

$$\mathbb{E}_{X=x}\left[\mathbb{E}_{\mathcal{D}}[(\mathbb{E}_{\mathcal{D}}[\widehat{f}_{\mathcal{D}}(x)] - \widehat{f}_{\mathcal{D}}(x))^2]\right] = \frac{\sigma^2}{n}\mathbb{E}_X[\mathrm{Trace}(\Sigma^{-1}XX^T)] = \frac{d\sigma^2}{n}$$

Example: Linear LS $\quad \mathbf{Y} = \mathbf{X}w + \epsilon$

$$\text{if} \quad y_i = x_i^T w + \epsilon_i \quad \text{and} \quad \epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$$

$$\widehat{w}_{MLE} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} = w + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\epsilon$$

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

$$\widehat{f}_{\mathcal{D}}(x) = \widehat{w}^T x = w^T x + \epsilon^T \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} x$$

$$\underbrace{\mathbb{E}_{XY}[(Y - \eta(x))^2 | X = x]}_{\textbf{irreducible error}} = \sigma^2 \qquad \underbrace{(\eta(x) - \mathbb{E}_{\mathcal{D}}[\widehat{f}_{\mathcal{D}}(x)])^2}_{\textbf{biased squared}} = 0$$

$$\mathbb{E}_{X=x}\Big[\underbrace{\mathbb{E}_{\mathcal{D}}[(\mathbb{E}_{\mathcal{D}}[\widehat{f}_{\mathcal{D}}(x)] - \widehat{f}_{\mathcal{D}}(x))^2]}_{\textbf{variance}}\Big] = \frac{d\sigma^2}{n}$$