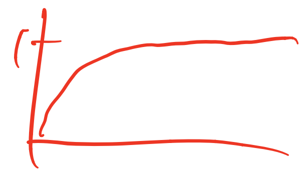
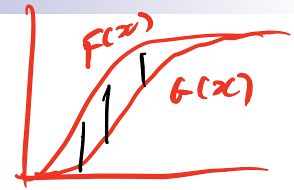
Accouncements

new $\{1-\frac{1}{n}\}$





HW0 Due Thursday at 11:59 PM

You should turn in a PDF and a python file(s)

$\sup_x |F(x) - G(x)|$

Figure for problem 9 should be in the PDF

Please do **not** zip these files and submit (unless there are >5 files)

Fix $x \in [0,1]$, $\mathbb{1}\{A\}=1$ if A is true, 0 otherwise

U is random variable, uniformly distributed on $[0,1]$

1. $\mathbb{E}[\mathbb{1}\{U \le x\}] = \int_0^1 \mathbb{1}\{u \le x\} \cdot 1\, du = \int_0^x 1 \cdot 1\, du = x$
2. $\mathbb{E}[\mathbb{1}\{U \le x\}^2] = \int_0^1 \mathbb{1}\{u \le x\}^2 \cdot 1\, du = \int_0^x 1 \cdot 1\, du = x$
3. $\mathbb{E}[U] = \int_0^1 u \cdot 1\, du = \frac{1}{2}u^2 \Big|_0^1 = \frac{1}{2}$
4. $\mathbb{E}[U^p]$ for $p > 0$

$$= \int_0^1 u^p \cdot 1 \, du = \frac{u^{p+1}}{p+1}\Big|_0^1 = \frac{1}{p+1}$$

Bayesian Methods

Machine Learning – CSE546
Kevin Jamieson
University of Washington

September 28, 2017

MLE Recap - coin flips

- **Data**: sequence *D= (HHTHT…)*, **k heads** out of **n flips**
- **Hypothesis:** P(Heads) = θ, P(Tails) = 1-θ

$$P(\mathcal{D}|\theta) = \theta^k (1-\theta)^{n-k}$$

- Maximum likelihood estimation (MLE): Choose θ that maximizes the probability of observed data:

$$\begin{aligned}\widehat{\theta}_{MLE} &= \arg\max_{\theta} \; P(\mathcal{D}|\theta) \\ &= \arg\max_{\theta} \; \log P(\mathcal{D}|\theta)\end{aligned}$$

$$\widehat{\theta}_{MLE} = \frac{k}{n}$$

MLE Recap - Gaussians

- MLE:

$$\log P(\mathcal{D}|\mu,\sigma) = -n\log(\sigma\sqrt{2\pi}) - \sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma^2}$$

$$\widehat{\mu}_{MLE} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \widehat{\sigma^2}_{MLE} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \widehat{\mu}_{MLE})^2$$

- MLE for the variance of a Gaussian is **biased**

$$\mathbb{E}[\widehat{\sigma^2}_{MLE}] \neq \sigma^2$$

  - Unbiased variance estimator:

$$\widehat{\sigma^2}_{unbiased} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \widehat{\mu}_{MLE})^2$$

MLE Recap

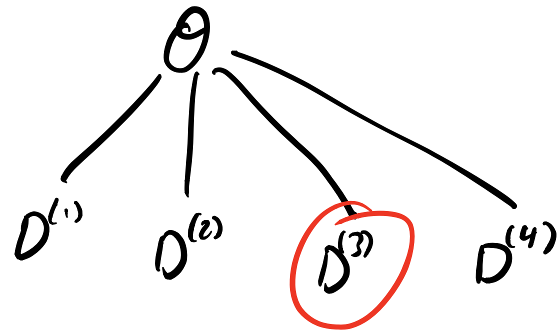
- Learning is…
  - Collect some data
    - E.g., coin flips
  - Choose a hypothesis class or model
    - E.g., binomial
  - Choose a loss function
    - E.g., data likelihood
  - Choose an optimization procedure
    - E.g., set derivative to zero to obtain MLE
  - Justifying the accuracy of the estimate
    - E.g., Hoeffding's inequality

What about prior

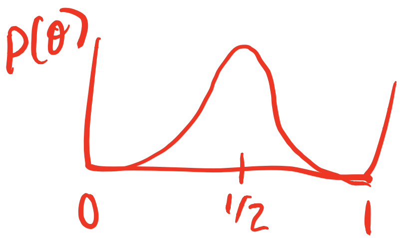
- *Billionaire*: Wait, I know that the coin is “close” to 50-50. What can you do for me now?
- **You say: I can learn it the Bayesian way…**

Bayesian vs Frequentist

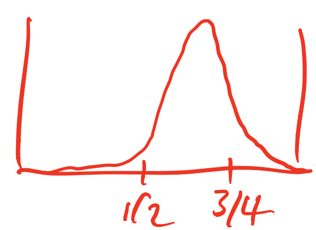
- Data: $\mathcal{D}$ Estimator: $\widehat{\theta} = t(\mathcal{D})$ loss: $\ell(t(\mathcal{D}), \theta)$
- Frequentists treat unknown θ ***as fixed*** and the data D ***as random***.



- Bayesian treat the data D **as fixed** and the unknown θ ***as random***



→ Observe $D = \frac{75}{100} \frac{\text{heads}}{\text{flips}}$ →



Bayesian Learning

- Use Bayes rule:

$$P(\theta \mid \mathcal{D}) = \frac{P(\mathcal{D} \mid \theta)P(\theta)}{P(\mathcal{D})}$$

- Or equivalently:

$$P(\theta \mid \mathcal{D}) \propto P(\mathcal{D} \mid \theta)P(\theta)$$

Bayesian Learning for Coins

$$P(\theta \mid \mathcal{D}) \;\propto\; P(\mathcal{D} \mid \theta)P(\theta)$$

- Likelihood function is simply Binomial:

$$P(\mathcal{D} \mid \theta) = \theta^{\alpha_H}(1-\theta)^{\alpha_T}$$

- What about prior?
  - Represent expert knowledge
- Conjugate priors:
  - Closed-form representation of posterior
  - **For Binomial, conjugate prior is Beta distribution**

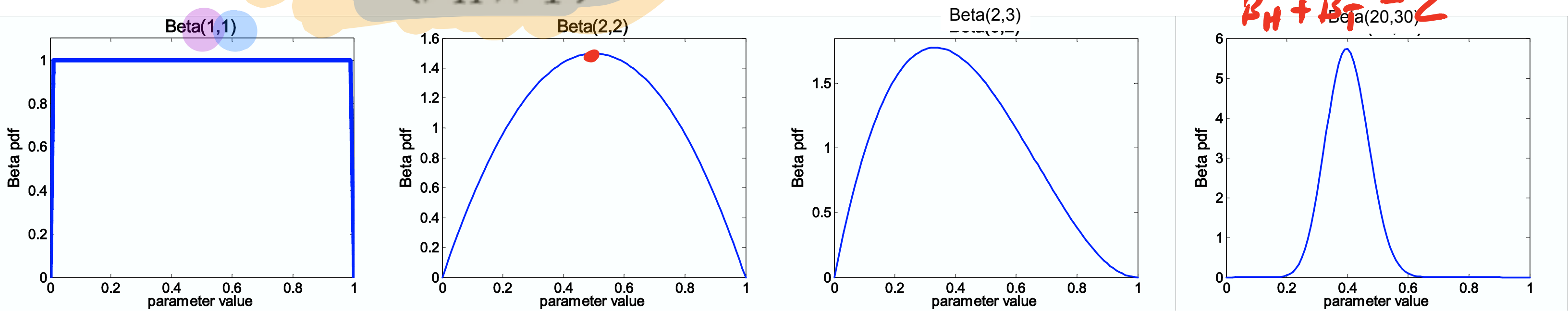
$p(D|\theta) = \theta^{\alpha_H}(1-\theta)^{\alpha_T}$ $\quad \hat{\theta}_{MLE} = \arg\max_\theta p(D|\theta) = \frac{\alpha_H}{\alpha_H+\alpha_T}$

Beta prior distribution – P(θ)

$$P(\theta) = \frac{\theta^{\beta_H-1}(1-\theta)^{\beta_T-1}}{B(\beta_H,\beta_T)} \sim Beta(\beta_H,\beta_T)$$

Mean: $= \frac{\beta_H}{\beta_H+\beta_T}$

Mode: $\frac{\beta_H-1}{\beta_H+\beta_T-2}$



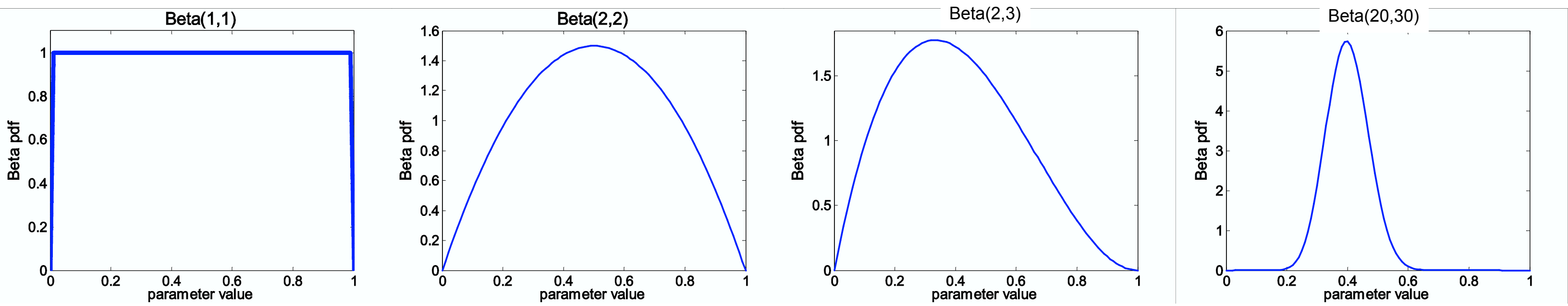
- Likelihood function: $P(\mathcal{D} \mid \theta) = \theta^{\alpha_H}(1-\theta)^{\alpha_T}$
- Posterior: $P(\theta \mid \mathcal{D}) \propto P(\mathcal{D} \mid \theta)P(\theta)$

Posterior distribution

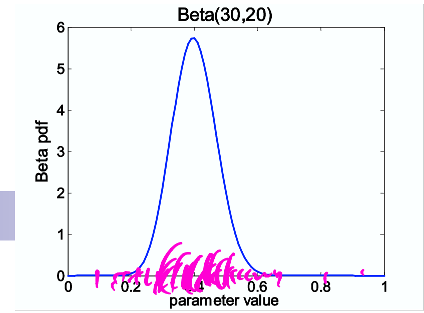
- Prior: $Beta(\beta_H, \beta_T)$
- Data: α_H heads and α_T tails

- Posterior distribution:

$$P(\theta \mid \mathcal{D}) = Beta(\beta_H + \alpha_H, \beta_T + \alpha_T)$$



Using Bayesian posterior



- Posterior distribution:

$$P(\theta \mid \mathcal{D}) \sim Beta(\beta_H + \alpha_H, \beta_T + \alpha_T)$$

- Bayesian inference:
  - No longer single parameter:

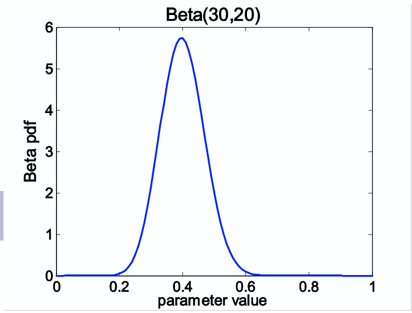
$$\theta_1, \theta_2, \dots, \theta_N \sim P(\theta \mid D)$$

$$\mathbb{E}[f(\theta) \mid D] \quad \cancel{E[f(\theta)]} = \int_0^1 f(\theta) P(\theta \mid \mathcal{D}) d\theta$$

$$\approx \frac{1}{N} \sum_{i=1}^{N} f(\theta_i)$$

  - Integral is often hard to compute

MAP: Maximum a posteriori approximation



$$P(\theta \mid \mathcal{D}) \sim Beta(\beta_H + \alpha_H, \beta_T + \alpha_T)$$

$$E[f(\theta)] = \int_0^1 f(\theta) P(\theta \mid \mathcal{D}) d\theta$$

- As more data is observed, Beta is more certain

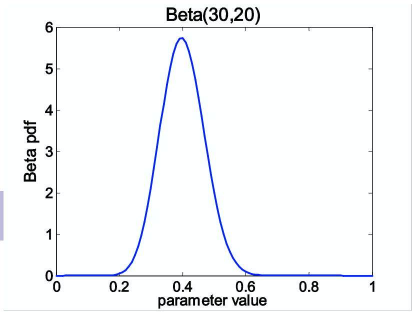
- MAP: use most likely parameter:

$$\hat{\theta}_{MAP} = \arg\max_{\theta} P(\theta \mid \mathcal{D}) \qquad E[f(\theta)] \cancel{\approx} f(\hat{\theta})$$

$$= \arg\max_{\theta} P(D|\theta)P(\theta)$$

$$\hat{\theta}_{MLE} = \arg\max_{\theta} P(D|\theta)$$

MAP for Beta distribution

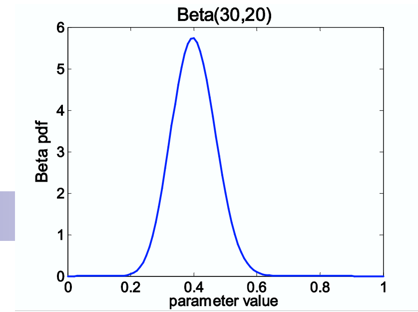


$$P(\theta \mid \mathcal{D}) = \frac{\theta^{\beta_H+\alpha_H-1}(1-\theta)^{\beta_T+\alpha_T-1}}{B(\beta_H+\alpha_H, \beta_T+\alpha_T)} \sim Beta(\beta_H+\alpha_H, \beta_T+\alpha_T)$$

- MAP: use most likely parameter:

$$\widehat{\theta}_{MAP} = \arg\max_{\theta} P(\theta \mid \mathcal{D}) = \frac{\alpha_H + \beta_H - 1}{\alpha_H + \beta_H + \alpha_T + \beta_T - 2}$$

MAP for Beta distribution



$$P(\theta \mid \mathcal{D}) = \frac{\theta^{\beta_H+\alpha_H-1}(1-\theta)^{\beta_T+\alpha_T-1}}{B(\beta_H+\alpha_H, \beta_T+\alpha_T)} \sim Beta(\beta_H+\alpha_H, \beta_T+\alpha_T)$$

- MAP: use most likely parameter:

$$\widehat{\theta} = \arg\max_{\theta} P(\theta \mid \mathcal{D}) = \frac{\beta_H + \alpha_H - 1}{\beta_H + \beta_T + \alpha_H + \alpha_T - 2}$$

- Beta prior equivalent to extra coin flips
- As $\cancel{N}\, M \rightarrow 1$, prior is "forgotten"
- **But, for small sample size, prior is important!**

Recap for Bayesian learning

- Learning is…
  - Collect some data
    - E.g., coin flips
  - Choose a hypothesis class or model
    - E.g., binomial and prior based on expert knowledge
  - Choose a loss function
    - E.g., parameter posterior likelihood
  - Choose an optimization procedure
    - E.g., set derivative to zero to obtain MAP
  - Justifying the accuracy of the estimate
    - E.g., If the model is correct, you are doing best possible

Recap for Bayesian learning

Bayesians are optimists:

- "If we model it correctly, we output most likely answer"
- Assumes one can accurately model:
  - Observations and link to unknown parameter θ: $p(x|\theta)$
  - Distribution, structure of unknown θ: $p(\theta)$

Frequentist are pessimists:

- "All models are wrong, prove to me your estimate is good"
- Makes very few assumptions, e.g. $\mathbb{E}[X^2] < \infty$ and constructs an estimator (e.g., median of means of disjoint subsets of data)
- Prove guarantee $\mathbb{E}[(\theta - \widehat{\theta})^2] \leq \epsilon$ under hypothetical true θ's

Linear Regression

Machine Learning – CSE546
Kevin Jamieson
University of Washington

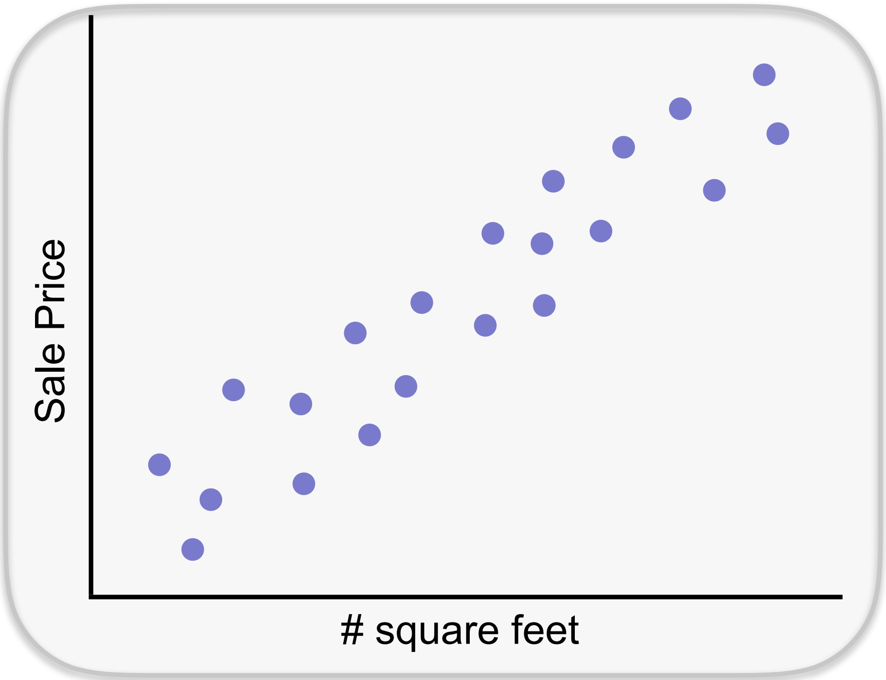
Oct 3, 2017

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:

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

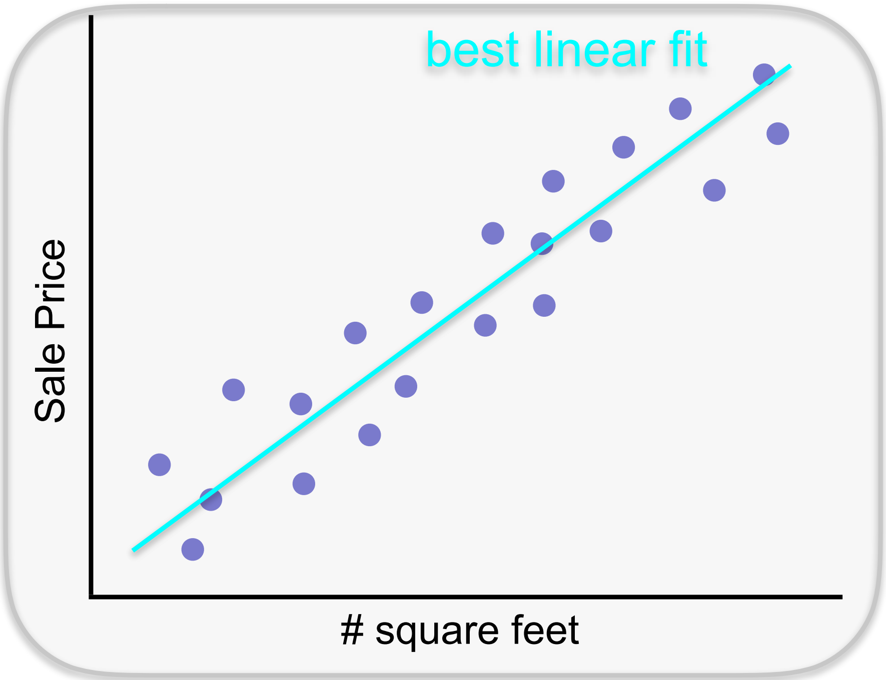
$$x_i \in \mathbb{R}^d$$
$$y_i \in \mathbb{R}$$

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:

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

Hypothesis: linear

$$y_i \approx x_i^T w = \sum_{j=1}^{d} x_{i,j} w_j$$

Loss: least squares

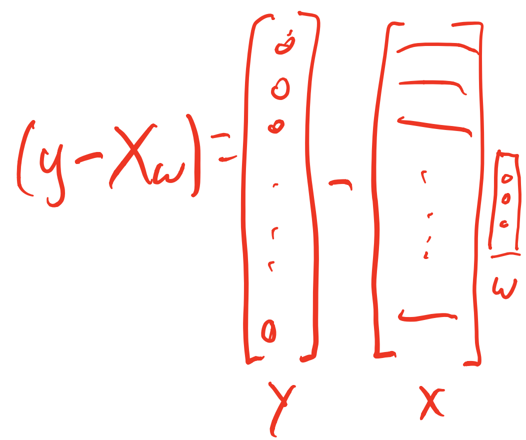
$$\min_w \sum_{i=1}^n \left(y_i - x_i^T w\right)^2$$

The regression problem in matrix notation

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

$$= \arg\min_w \underbrace{(\mathbf{y} - \mathbf{X}w)^T}_{1 \times n} \underbrace{(\mathbf{y} - \mathbf{X}w)}_{n \times 1} = \|y - Xw\|_2^2$$

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



$\|u\|_p = \left(\sum_{i=1}^n |u_i|^p\right)^{1/p}$ $u \in \mathbb{R}^n$ $\|u\|_2 = \sqrt{\sum_{i=1}^n u_i^2}$

The regression problem in matrix notation

$$\begin{aligned}\widehat{w}_{LS} &= \arg\min_w ||\mathbf{y} - \mathbf{X}w||_2^2 \\ &= \arg\min_w \underbrace{(\mathbf{y} - \mathbf{X}w)^T(\mathbf{y} - \mathbf{X}w)}_{= J(w)}\end{aligned}$$

$$\nabla_w J(w) = \nabla_w \left(y^T y - y^T X w - (Xw)^T y + w^T X^T X w \right)$$

$$= 0 - X^T y - X^T y + 2 X^T X w$$

$$= -2X^T y + 2X^T X w = 0 \qquad X^T X \hat{w}_{LS} = X^T y$$

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

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?

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

Dealing with an offset

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

Dealing with an offset

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

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

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

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

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

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 $\quad \epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$

$$P(y|x, w, \sigma) =$$

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

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

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

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

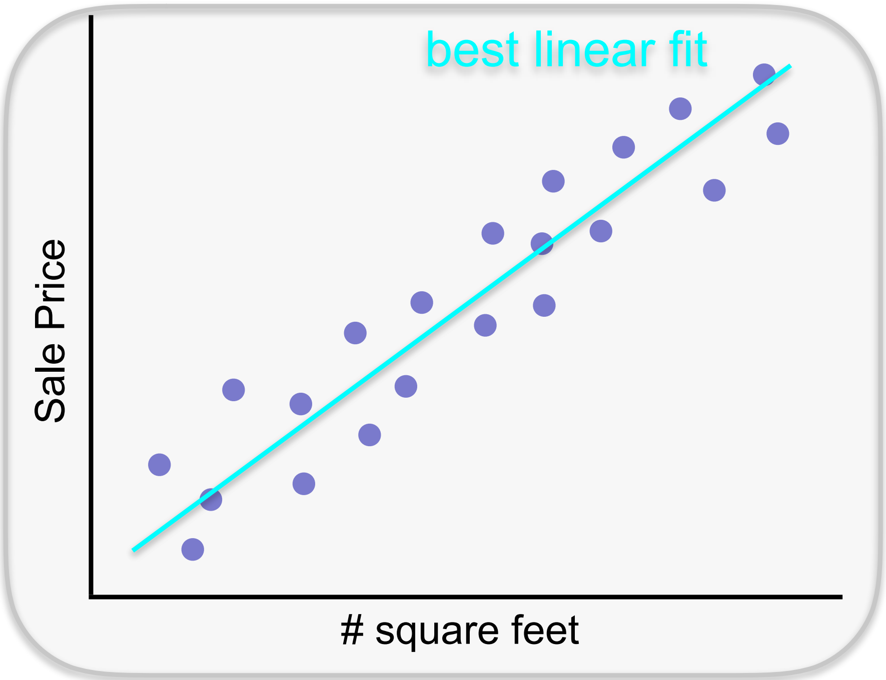
$$\boxed{\widehat{w}_{LS} = \widehat{w}_{MLE} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}}$$

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

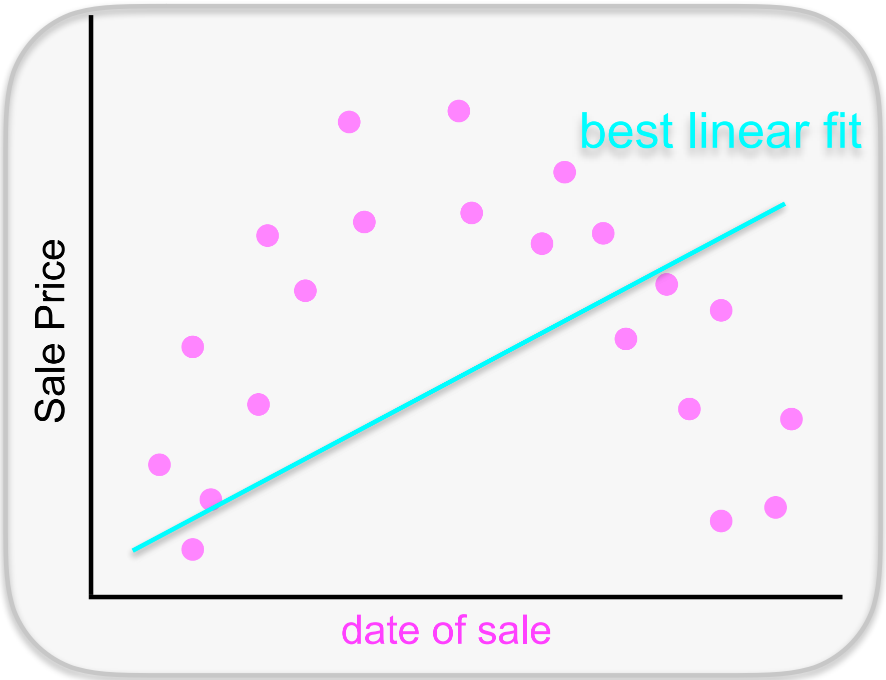
$$\min_w \sum_{i=1}^n \left(y_i - x_i^T w\right)^2$$

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

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

Training Data:

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

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

Hypothesis: linear

$$y_i \approx x_i^T w$$

Loss: least squares

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

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 $\{u_j\}_{j=1}^p \subset \mathbb{R}^d$

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

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

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

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~~

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

~~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

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

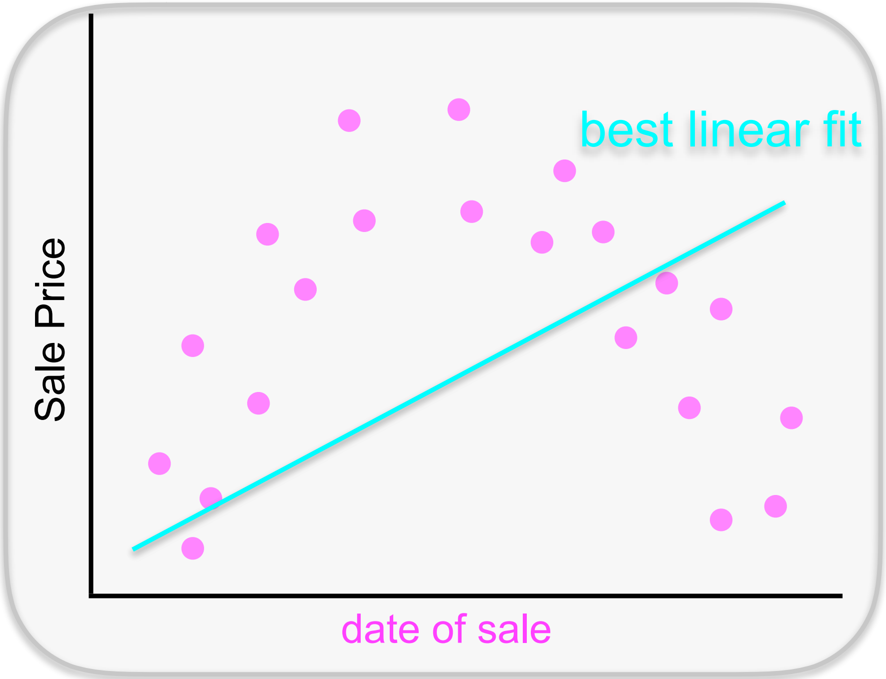
Loss: least squares

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

The regression problem

Training Data:

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



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

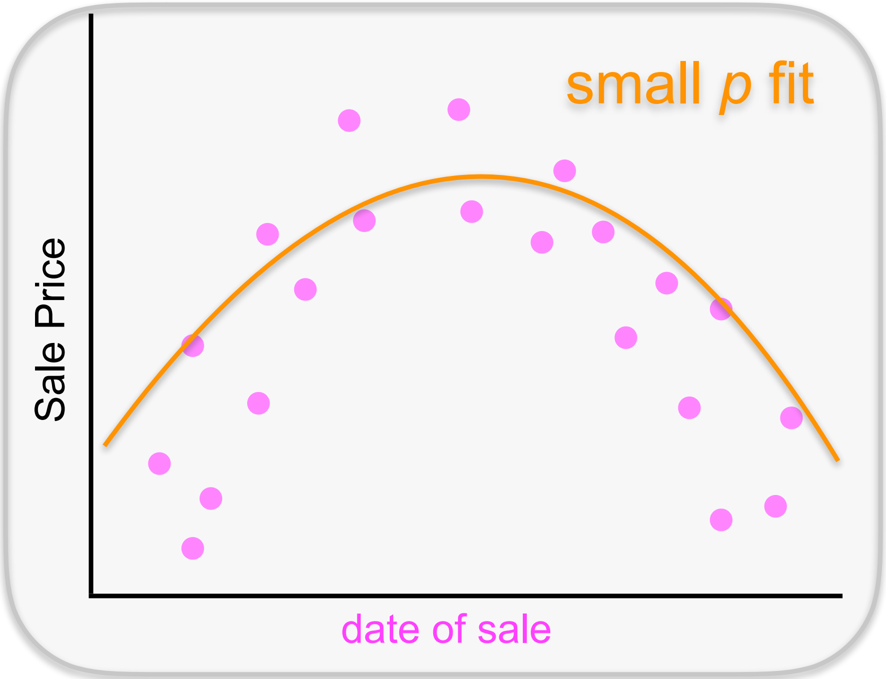
$$\min_w \sum_{i=1}^n \left(y_i - h(x_i)^T w\right)^2$$

The regression problem

Training Data:

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

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



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

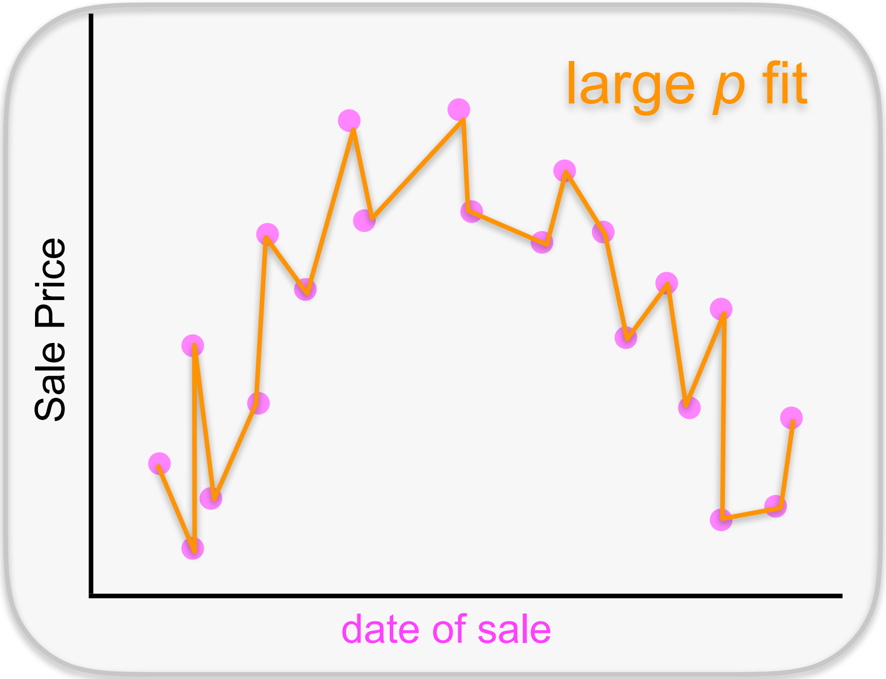
$$\min_w \sum_{i=1}^n \left(y_i - h(x_i)^T w\right)^2$$

The regression problem

Training Data:

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

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



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

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

What's going on here?