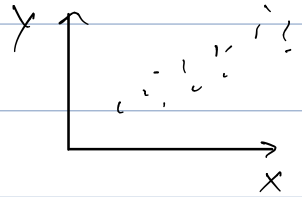
Day 1

9/7/18

<u>Goal of supervized machine learning:</u>

use labeled example data to make and interpret predictions



House sales	X: sq feet	Y: price
Advertisement	X: spending	Y: sales
HR	X: age	Y: salar

<u>Process of supervised ML:</u>

Populaton $\xrightarrow[(x_i, y_i)]{\text{data collection}}$ Data

$Y = \theta_0 + \theta_1 X + \varepsilon$

$y_i = \hat{\theta}_0 + \hat{\theta}_1 x_i + e_i$

conclusions
- prediction
- interpretation

<u>Components:</u> x_i - input, $x_i \in \mathcal{X} = \mathbb{R}^p$

y_i - output; target, $y_i \in \mathcal{Y} = \mathbb{R}$

$i = 1, \ldots, N$

(x_i, y_i) - training example

$\{(x_i, y_i)\}_{i=1}^N$ - training set

<u>Goal:</u> develop function $h(x): \mathcal{X} \to \mathcal{Y}$

$h(x)$ - hypothesis, model

$h_\theta(x)$ - hypothesis depends on parameters θ

<u>Process:</u>
(1) define problem space
(2) collect data
(3) specify model / hypothesis
(4) develop learning algorithm (i.e., find values of params)
(5) make predictions

Terminology $h_\theta(x): \mathcal{X} \to \mathcal{Y}$ - supervized ML
$\mathcal{Y} = \mathbb{R} \to$ regression
$\mathcal{Y} =$ set of discrete values $\}$ - classification
patterns of $\mathcal{X}$ - unsupervized ML

Simple (one-variable) linear regression

Step (3): $h_\theta(x) = \underbrace{\theta_0 + \theta_1}_{\substack{\text{parameters}\\ \text{weights}\\ \text{constants}}} x = \theta_0 \cdot \underbrace{1 + \theta_1 \cdot x}_{\substack{\text{fixed}\\ \text{basis}}}$

Vector notation: $h_\theta(x) = [1 \ x] \begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix} = \underset{1\times 2}{x^T} \underset{2\times 1}{\theta}$

Step (4): to estimate (i.e., determine from data) the values of θ_0 and θ_1, minimize cost function (= loss function)

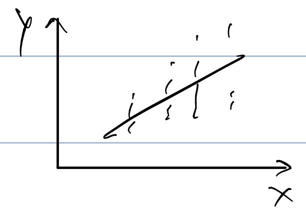
$J(\theta) = \sum_{i=1}^{N} (y_i - h_\theta(x_i))^2$ - squared loss
(many other loss functions are possible)

Params $\hat{\theta} = \arg\min_\theta J(\theta)$ - least squares estimates

Prediction: $\hat{y}_i = \hat{\theta}_0 + \hat{\theta}_1 x_i$

Probabilistic interpretation

Randomness in the data: consider probability distribution of Y given values of X



$P\{Y|X\}$ - class of such models are called discriminative models

Special case: Normal linear regression

Assume: $Y_i | X_i \overset{indep}{\sim} \mathcal{N}(\theta_0 + \theta_1 X_i, \sigma^2)$

Equivalently: $Y_i | X_i = \theta_0 + \theta_1 X_i + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$

independent and identically distributed

Decomposition of the overall variation:

$$E\{(Y - \hat{Y})^2\} = E\{(h_\theta(X) + \varepsilon - \hat{h}_\theta(X))^2\} =$$

$$= E\{(h_\theta(X) - \hat{h}_\theta(X))^2 + 2E\{(h_\theta(X) - \hat{h}_\theta(X))\varepsilon\} + E\{\varepsilon^2\}$$

$$= E\{(h_\theta(X) - \hat{h}_\theta(X))^2\} + \sigma^2$$

reducible error — irreducible error

Parameter estimation: maximum likelihood

$$f(Y_i | X_i, \theta) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2\sigma^2}(Y_i - \theta_0 - \theta_1 X_i)^2}$$

Probability distribution: $f(\text{data} | \text{params})$

Likelihood: $f(\text{params} | \text{data})$

Params that maximize the likelihood are maximum likelihood estimates

Bivariate Normal distributions: regression vs correlation

$$f(Y_1, Y_2) = \frac{1}{2\pi \sigma_1 \sigma_2 \sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(Y_1 - \mu_1)^2}{\sigma_1^2} - 2\rho\left(\frac{Y_1 - \mu_1}{\sigma_1}\right)\left(\frac{Y_2 - \mu_2}{\sigma_2}\right) + \frac{(Y_2 - \mu_2)^2}{\sigma_2^2}\right]\right\}$$

where ρ is the coef. of correlation

$$\rho = \frac{\text{cov}(Y_1, Y_2)}{\sigma_1 \sigma_2} = \frac{E\{(Y_1 - \mu_1)(Y_2 - \mu_2)\}}{\sigma_1 \sigma_2}$$

Can show: $Y_1 | Y_2$ is Normally distributed

$$E\{Y_1 | Y_2\} = \underbrace{\left(\mu_1 - \mu_2 \rho \frac{\sigma_1}{\sigma_2}\right)}_{\theta_0} + \underbrace{\rho \frac{\sigma_1}{\sigma_2}}_{\theta_1} Y_2$$

$$\text{Var}\{Y_1 | Y_2\} = \sigma_1^2 (1 - \rho^2)$$

- variance of the conditional distr. ↓ if correlation ↑ -