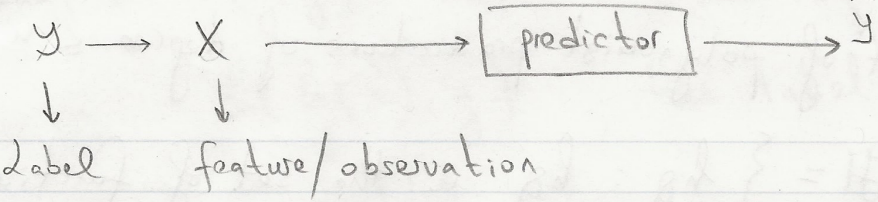
Lecture 5:

Inference Problems:



Example: X: avacodo or more precisely the features of an avacodo, e.g. the color of an avacodo, $X \in \mathcal{X} = [0,1]$.

y: goodness of an avacado $y \in \mathcal{Y} = [0,1]$
goodness of an avacodo $y \in \mathcal{Y} = \{0,1\}$

Example: X: Images $64 \times 64 = 4096$ pixels.

$\bar{X} \in \mathbb{R}^{64 \times 64}$.

y: Class of an image, e.g. $y \in \mathcal{Y} = \{0,1\}$ (cats, dogs)

Goal: Find a predictor $h: \mathcal{X} \to \mathcal{Y}$ that can accurately predict y from X.

Data-Driven Approach:

Data: $\{(x_i, y_i)\}_{i=1}^{n}$ E.g. n labeled avacados, n labeled images.

Hypothesis class:

$\mathcal{H}$: set of functions we can use to implement the predictor

$\mathcal{H} = \{ h_\theta : h_\theta(\bar{x}) = \bar{\theta}^T \bar{x} = \sum_{i=1}^{d} \theta_i x_i \text{ where } \bar{\theta} \in \mathbb{R}^d, \bar{x} \in \mathbb{R}^d \}$

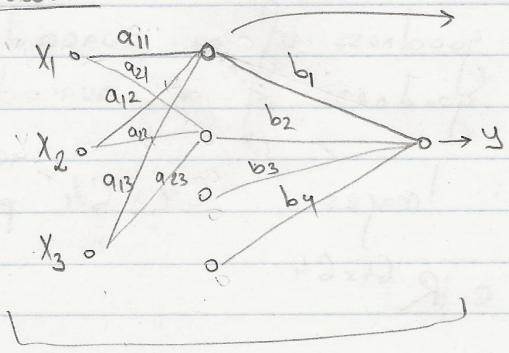
→ Linear Predictors.

Example: $\mathcal{H} = \{ h_\theta(x) = \theta_1 x^n + \theta_2 x^{n-1} + \dots \theta_n x + \theta_{n+1} \}$

(set of polynomial predictors of degree $\leq n$.

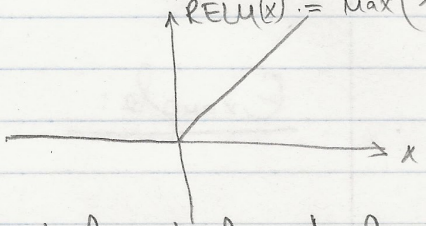
Example $\mathcal{H} = \{ h_\theta : h_\theta$ is the set of functions we can implement with a neural network with a given structure and coefficients $\theta \in \mathbb{R}^d \}$

Neural network:



$h_1 = \text{ReLU}(a_{11}x_1 + a_{12}x_2 + a_{13}x_3)$

$\text{RELU}(x) = \max(0, x)$



$y = b_1 h_1 + b_2 h_2 + b_3 h_3 + b_4 h_4$

$f : \mathbb{R}^3 \to \mathbb{R}$

Loss function: $\ell : Y \times Y \to \mathbb{R}^+$ non-negative.

$\ell(\hat{y}, y)$: takes two labels and quantifies how different the two labels are.

E.g. $y \subseteq \mathbb{R}$. regression problems.

$\ell(\hat{y}, y) = (\hat{y} - y)^2$. squared loss.

E.g. $y \in \{1, 2, \dots k\}$ discrete set classification problems with k classes.

$\ell(\hat{y}, y) = \begin{cases} 1 & \text{if } y \neq \hat{y} \\ 0 & \text{o/w} \end{cases} = 1_{\{\hat{y} \neq y\}}(\hat{y}, y)$

0-1 Loss.

The loss of a predictor $h \in \mathcal{H}$ on a given sample (x, y) will be given by

$$\ell(h(x), y)$$

$h(x)$: the label predicted by h for x; y: true label of x

Empirical Loss of a predictor h: (the average loss of a predictor h on our sample set $\{(x_i, y_i)\}_{i=1}^{n}$)

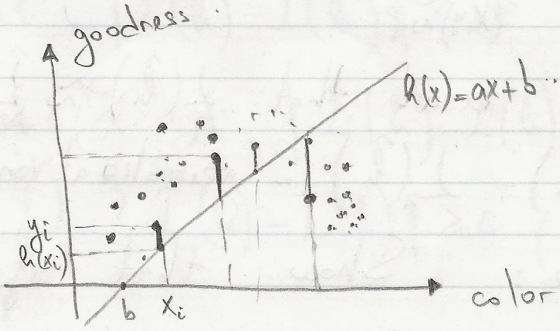
$$\mathcal{L}_n(h) = \frac{1}{n} \sum_{i=1}^{n} \ell(h(x_i), y_i)$$

Empirical Risk Minimization (ERM):

Find $h \in \mathcal{H}$ that has the smallest $\mathcal{L}_n(h)$

$$\hat{h} = \underset{h \in \mathcal{H}}{\operatorname{argmin}} \; \mathcal{L}_n(h) = \underset{h \in \mathcal{H}}{\operatorname{argmin}} \; \frac{1}{n} \sum_{i=1}^{n} \ell(h(x_i), y_i)$$

Ex:



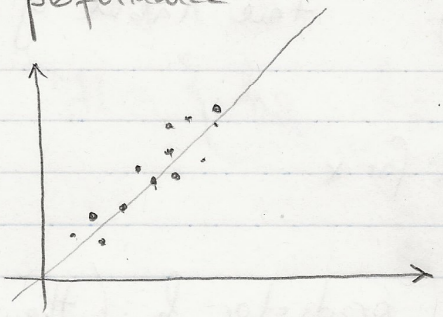
$$\mathcal{H} = \{ h_{(a,b)} : h_{(a,b)} = ax + b \}$$

$$\ell(h(x_i), y_i) = (h(x_i) - y_i)^2 = (ax_i + b - y_i)^2$$

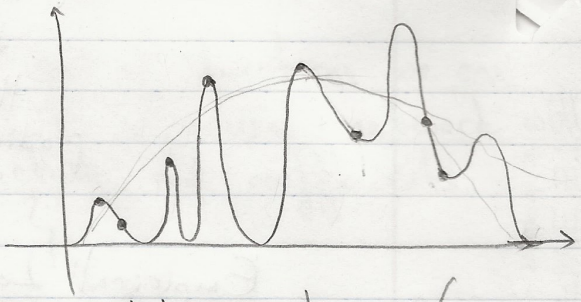
$$\min_{h \in \mathcal{H}} \mathcal{L}_n(h) = \min_{a,b} \frac{1}{n} \sum_{i=1}^{n} (ax_i + b - y_i)^2$$

Linear Least Squares Regression.

How do we know that the performance of a chosen predictor $\hat{h}$ on the data set is representative of its performance in the real world / population?



The data may not be representative



Memorization / Poor generalization. Overfitting.

Data generation model:

$$\{(x_i, y_i)\}_{i=1}^{n} \overset{i.i.d}{\sim} P \quad (P \text{ is the true (unknown) distribution from which nature generates } (X,Y))$$

Population loss of a predictor h:

$$L(h) = \mathop{\mathbb{E}}_{(X,Y)\sim P}\left[\ell(h(x), y)\right] = \int \ell(h(x), y) f_{X,Y}(x,y)\, dx$$

How can we ensure that $L(\hat{h}) \approx L_n(\hat{h})$:

$L(\hat{h}) - L_n(\hat{h})$ = generalization error.

Uniform convergence: Show that if we take n i.i.d. samples from P and evaluate $L_n(h)$ for all $h \in \mathcal{H}$, then $L_n(h) \approx L(h)$ for all $h \in \mathcal{H}$ (regardless of P). Then no matter what $\hat{h} \in \mathcal{H}$ is chosen by the ERM algorithm, $L_n(\hat{h}) \approx L(\hat{h})$, i.e. the ERM solution will generalize well to the real world.