Lec 7 Math 390.4 2/21/18

New idea for a learning algorithm...

$$g = \mathcal{A}(\mathbb{D}, \mathcal{H}) \quad . \quad \text{Predict via } \hat{y}^* = g(\vec{x}^*)$$

What if g found the "closest" $\vec{x}_i \in \mathbb{D}$ to $\vec{x}^*$

and returned $\hat{y}^* = y_i$? This closest $\vec{x}_i$ is called its neighbor.

Need to define "closest" via a distance function

$$d(\vec{x}_i, \vec{x}_k) = \underbrace{\|\vec{x}_i - \vec{x}_k\|_2^2}_{\text{Euclidean norm sqd.}} = (\vec{x}_i - \vec{x}_k)^T(\vec{x}_i - \vec{x}_k) = \sum_{j=1}^{p} (x_{ij} - x_{kj})^2$$

many others...

$$\left.\begin{array}{l} \mathcal{H} = \{ \, ? \, \} \\ \mathcal{A} = ? \end{array}\right\} \text{ difficult to define precisely}$$

$g(\vec{x}^*)$ ... all the work happens here...

Extension: find the K closest $\vec{x}_i$'s. Then $\hat{y} = \text{Mode}[y_{(1)}, \ldots, y_{(K)}]$

returns most likely class

K nearest neighbors or "KNN" algorithm.

Weaknesses? p large AKA "curse of dimensionality"

and not all x_j features are equally predictive!

Choices? K, d — they really matter!! Learning is not simple.

So far, we have been concerned with problems s.t. $\mathcal{Y} = \{0,1\}$. This is called "binary classification". If $\mathcal{Y} = \{0,1,\ldots,K\}$ where the response levels are nominal (i.e. no order), this is called "classification" or "multilevel classification".

What if $\mathcal{Y} \in \mathbb{R}$ or $\mathcal{Y} \in R \subset \mathbb{R}$? This is then called "regression".
Why? For a historical reason which we will get to.

Can the threshold, perceptron or SVM do regression? Not without serious adaptation. What do we do?? Null model? $\mathcal{H} = \{y_0 : y_0 \in \mathbb{R}\}$
$g = \bar{y}$ Avg. value

Recall $\mathcal{H} = \left\{ \mathbb{1}_{\underbrace{\vec{w} \cdot \vec{x}}_{\text{linear model}}} : \vec{w} \in \mathbb{R}^{p+1} \right\}$

Why the indicator function? To coerce the output $g \in \{0,1\}$ instead of $\mathbb{R}$.
Why not use just the linear model?

convention to index this at 0 and not call it b.

$$\mathcal{H} = \{\vec{w} \cdot \vec{x} : \vec{w} \in \mathbb{R}^{p+1}\} = \{w_0 + w_1 x_1 + \ldots + w_p x_p : w_0 \in \mathbb{R}, w_1 \in \mathbb{R}, \ldots, w_p \in \mathbb{R}\}$$

~~This is the most famous model period. The "linear regression model". For historical reasons we will denote the weights as $\vec{b}$, not $\vec{w}$.~~

$$\cancel{\mathcal{H} = \{\vec{b} \cdot \vec{x} : \vec{b} \in \mathbb{R}^{p+1}\} = \{b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_p x_p : b_0 \in \mathbb{R}, b_1 \in \mathbb{R}, \ldots, b_p \in \mathbb{R}\}}$$

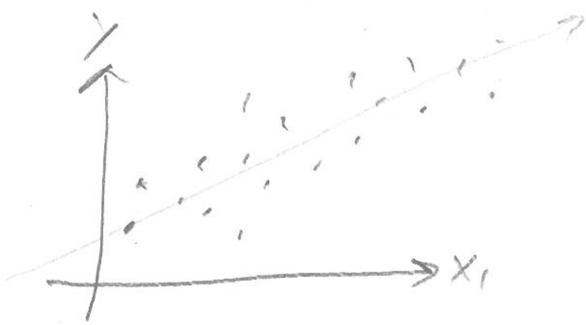
convention

Dim of param space: $p+1$

We can visualize this if $p = 1$

$\vec{w}$ is the linear coefficients

$$\mathcal{H} = \{w_0 + w_1 x_1 : w_0 \in \mathbb{R}, w_1 \in \mathbb{R}\}$$



$h^*(\vec{x}) = w_0^* + w_1^* x = \beta_0 + \beta_1 x$

the candidate in $\mathcal{H}$ that most closely resembles f.

special notation for the optimal/true coefficients

Review: why doesn't $h^* = y$ exactly?

$$y = h^*(\vec{x}) + \varepsilon = h^*(\vec{x}) + \underbrace{\overbrace{\left(t(\vec{z}) - f(\vec{x})\right)}^{\text{ignorance}} + \overbrace{\left(f(\vec{x}) - h^*(\vec{x})\right)}^{\text{misspecification of linear model}}}_{\varepsilon}$$

AKA "noise" or "errors"

h^* is inaccessible since we have to make an imperfect fit with finite data

$$y = g(\vec{x}) + \underset{\text{"residuals"}}{e} = g(\vec{x}) + \underbrace{\left(t(\vec{z}) - f(\vec{x})\right) + \left(f(\vec{x}) - h^*(\vec{x})\right)}_{\varepsilon} + \underbrace{\overbrace{\left(h^*(\vec{x}) - g(\vec{x})\right)}^{\text{estimation error}}}_{e - \varepsilon}$$

As $n \to \infty$ $g(\vec{x}) \to h^*(\vec{x})$ and $e - \varepsilon \to 0$ but $y \neq g(\vec{x})$ since the other two errors are still present.

Back to the linear model for $p=1$ (AKA "simple linear regression"). How to fit $\vec{w}$, the parameters? Need an $\mathcal{A}$.

First need a **loss function**, error function, objective function, cost function.

Recall before $\underset{x_i}{\overset{\text{min:}}{SSE}} := \sum_{i=1}^{n} \underbrace{\left(y_i - \hat{y}_i\right)}_{e_i}^2 = \sum e_i^2$ "sum of squared error"

$$= \sum_{i=1}^{n} \left(y_i - (w_0 + w_1 x_i)\right)^2 = \sum_{i=1}^{n} y_i^2 + w_0^2 + w_1^2 x_i^2 - 2y_i w_0 - 2y_i w_1 x_i + 2w_0 w_1 x_i$$

$\bar{y} = \frac{1}{n} \sum y_i$

$$= \sum y_i^2 + n w_0^2 + w_1^2 \sum x_i^2 - 2n\bar{y} w_0 - 2 w_1 \sum x_i y_i + 2 w_0 w_1 n \bar{x}$$

Choose w_0, w_1 to min. the above

$$\frac{\partial}{\partial w_0}\left[\ \right] = 2n w_0 - 2n\bar{y} + 2 w_1 n \bar{x} \overset{\text{set}}{=} 0 \Rightarrow \hat{w}_0 = \bar{y} - w_1 \bar{x}$$

$$\frac{\partial}{\partial w_1}\left[\quad\right] = \cancel{2}w_1 \Sigma x_i^2 - \cancel{2}\Sigma y_i x_i + \cancel{2}w_0 n\bar{x} = 0$$

$$= w_1 \Sigma x_i^2 - \Sigma y_i x_i + (\bar{y} - w_1\bar{x})\, n\bar{x} = 0$$

$$= w_1 \Sigma x_i^2 - \Sigma y_i x_i + n\bar{x}\bar{y} - w_1 n\bar{x}^2 = 0$$

$$= w_1\left(\Sigma x_i^2 - n\bar{x}^2\right) = \Sigma y_i x_i - n\bar{x}\bar{y}$$

$$\Rightarrow \hat{w}_1 = \frac{\Sigma y_i x_i - n\bar{x}\bar{y}}{\Sigma x_i^2 - n\bar{x}^2} = \frac{\cancel{(n-1)} s_{XY}}{\cancel{(n-1)} s_X^2} = r\frac{s_Y}{s_X} \Rightarrow \hat{w}_0 = \bar{y} - r\frac{s_Y}{s_X}\bar{x}$$

Statistics Note:

Note: $\rho := \mathrm{Corr}[X,Y] = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}}$ $\sigma_{XY}/\sigma_X\sigma_Y$ is the correlation between X, Y

$\sigma_{XY} := \mathrm{Cov}[X,Y] = E[(X-\mu_X)(Y-\mu_Y)]$ " " covariance " " X, Y

$s_{XY} := \frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{n-1}$ is the sample covariance. It estimates σ_{XY}

$$= \frac{\Sigma x_i y_i - \Sigma x_i \bar{y} - \Sigma y_i \bar{x} + n\bar{x}\bar{y}}{n-1}$$

$$= \frac{\Sigma x_i y_i - n\bar{x}\bar{y} \cancel{- n\bar{x}\bar{y} + n\bar{x}\bar{y}}}{n-1} = \frac{\Sigma x_i y_i - n\bar{x}\bar{y}}{n-1}$$

Note: $s_X^2 := \frac{1}{n-1}\Sigma(x_i-\bar{x})^2$ is the estimate of σ_X^2 (sample variance)

$$= \frac{1}{n-1}\left(\Sigma x_i^2 - 2\Sigma x_i\bar{x} + \Sigma\bar{x}^2\right)$$

$$= \frac{1}{n-1}\left(\Sigma x_i^2 - 2n\bar{x}^2 + n\bar{x}^2\right)$$

$$= \frac{1}{n-1}\left(\Sigma x_i^2 - n\bar{x}^2\right)$$

Note $r = \frac{s_{XY}}{s_X s_Y}$ is the sample correlation coefficient, the estimate of ρ.

$\hat{w}_0, \hat{w}_1$ have a special notation: $b_0, b_1 \Rightarrow g(x) = b_0 + b_1 x$

And to predict...

$\hat{y}^* = g(x^*) = b_0 + b_1 x^*$

Why is this called "least squares regression"?