Class Notes 10/28/24

- Admin Info:
  - lab 6 due today
  - lab 7 due next Mon (Nov 4)
  - final project proposal due next Fri (Nov 8)
    - Nov 8 - Dec 4 = work on project
    - Dec 9 & 11 = in class presentations
    - Dec 20 = github repos must be finalized
  - find a dataset, run an alg on the dataset, do a comparison, evaluate results

email:
- title & names
- dataset (n & p)
- algorithm you will develop
- scientific? you're answering
- way to evaluate, interpret, & visualize results
- references

8 min per pair & 4 min for questions & feedback

Project Code, Lab Notebook (README.md), presentation slides

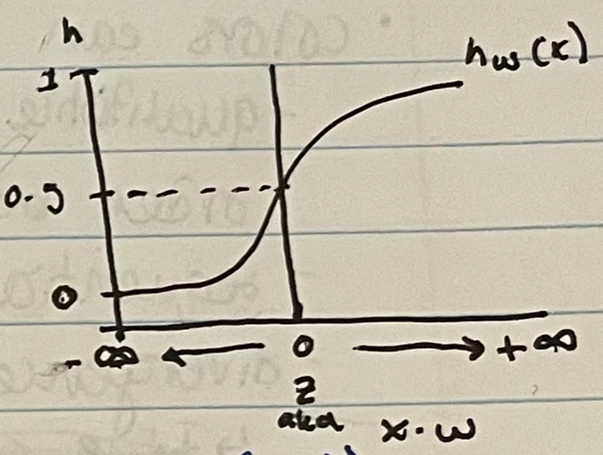
running doc [who was working, data, how long, what accomplished]

- SGD:

Hypothesis Function (Prediction):

- maps input value ($z = w \cdot x$) to probability btwn 0 & 1

$$h_w(x) = p(y = 1 \mid x) = \frac{1}{1 + e^{-w \cdot x}}$$

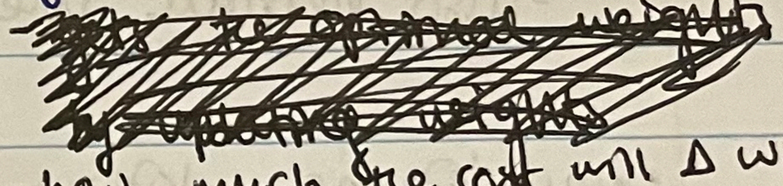


Cost Function (want to minimize):

$$J(w) = -\sum_{i=1}^{n} y_i \log h_w(x_i) + (1 - y_i) \log(1 - h_w(x_i))$$

Gradient of cost w/ respect to a single data point x_i:

$$\nabla J_{x_i}(w) = (h_w(x_i) - y_i) x_i$$



- how much the cost will Δ w/ a specific weight

- if $w \cdot x < 0 \rightarrow$ prediction (h) $< 0.5 \rightarrow$ predict $y = 0$
- if $w \cdot x \geq 0 \rightarrow$ $h \geq 0.5 \rightarrow$ predict $y = 1$
- SGD for Logistic Regression (binary classification):

set $\vec{w} = \vec{0}$

while cost $J(\vec{w})$ is still changing:

shuffle data points

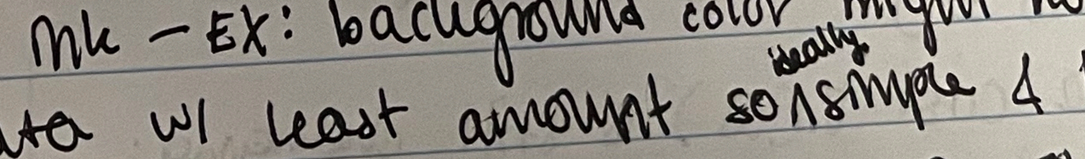
for $i = 1, \ldots, n$:

$\vec{w} \leftarrow \vec{w} - \alpha \nabla J_{\vec{x_i}}(\vec{w})$ ← derivative of $J(\vec{w})$ wrt x_i

Store $J(\vec{w})$

updates the weights based on the gradient of the cost func J wrt x_i

- visualizations are important and want to prevent being:
  - ugly = figure has aesthetic problems but otherwise clear & informative
  - bad = problems related to perception (unclear, confusing, deceiving)
  - wrong = incorrect, mathematical problems
- data types:
  - continuous vs. discrete
  - can have ordered & unordered qualitative / categorical data
- aesthetics in data visualization:
  - based on data types can Δ position, shape, size, color, line width, & line type
  - continuous: best option is position on a common scale |:: but can also use a color or area spectrum
  - discrete: could group different shapes or colors
  - ↳ can use multiple techniques at the same time - Ex: graph may use same scale for lines and color
- colors can be helpful to visualize data:
  - qualitative: use color to distinguish groups, don't want color to show an order between groups
  - sequential: use color gradient to show ordering of data
  - divergence: use sequential scales based on middle point
    - ↳ ~~blue color gra~~



- be aware of color-blind friendly colors
  - also helpful when print in black & white

- overplotting = plotting points on top of eachother
  - can use just the outline
  - jitter = seperate points
  - can use transparency to see where points overlap
- don't want to use excess ink - Ex: background color might not be helpful
  - want to showcase data w/ least amount of ideally simple & legible
  - can use legend or directly label lines on a graph
- double encoding = coding the same thing twice (Ex: color & scale to show profit) - can be distracting
  - can sometimes be helpful - Ex: color a group the same & use shapes to outline

error bar = best practice to say where error bar is (redundant, denoting)