- Converting features
    - Step 1:
        - Sort based on feature values
    - Step 2:
        - If different labels have the same feature values-collapse and label as none
    - Step three
        - Whenever the label changes, average the adjacent values when the feature value changes i.e, going from yes to no in handout example
- Logistic Regression
    - **Slide 10**: If using linear regression, y has to be real valued, how would you encode y?
        - stroke=0, drug overdose=1 and epileptic seizure would=2
        - This encoding would assume an ordering that is not necessarily there.
        - Linear regression is a bad choice because it assumes that multiple things cannot be true at the same time. Also we want the output to be the probability of one of the conditions being true, so we want the range to be [0,1], but the range of a linear function is infinite
        - If there is a decision boundary a point at which the prediction will change, we cannot obtain an accurate prediction with more than 2 features
    - logistic regression is used for binary classification
    - Z in the sigmoid function is replaced with -w*x
    - To perform classification given logistic regression model, if $w*x \geq 0$, the predicted value for y will be 1 and if $w*x<0$, y will be 0 since the output of $h_w(x)$ means the output of the probability so if it is less than 50% we predict 0 and > 50% we predict 1
    - We use this because decision trees overfit to small data since they overfit, also, with larger decision trees parsing becomes difficult. Logistic regression minimizes this by simplifying classification
- Cost function and SGD for logistic regression
    - Cost function for logistic regression measures model performance by using likelihood
    - We want the value of the likelihood function to be high since this means the model is performing well, high likelihood for the probabilities aligning with given points
    - To find the cost, we find the negative log of the likelihood. We take the log to add together the probabilities instead of multiplying them
    - Image 2
    - For a single data point, $h_w(x)$ can be replaced with y for the cases to calculate cost
- Cost function as Cross Entropy
    - Cross entropy is a measure of the difference between the two probabilities measure, it also measures the difference between between true Y and predicted Y, and we want entropy to be low meaning the difference is low