CHE 599: machine learning primer

what is machine learning?

machine learning is the study of algorithms and statistical/mathematical models that allow computers to perform tasks without being programmed with explicit instructions. (i.e. if … then … else …)

- instead, we construct a mathematical model & tune it using training examples (data)
- this model is then used to make predictions / decisions or to gain insights into the structure of our data set

supervised learning

distinction: training examples are labeled
goal: learn a function that maps an input $\vec{x}$ (feature vector) to an output $\vec{y}$ (target or response vector, dependent variable) $\vec{x} \mapsto \vec{y}$, from training examples
training examples: data $D := \{(\mathbf{x}_1, \mathbf{y}_1), ..., (\mathbf{x}_i, \mathbf{y}_i), ..., (\mathbf{x}_n, \mathbf{y}_n)\}$ example input/output pairs

(n examples; $\mathbf{x}_i$: feature vector for example i; $\mathbf{y}_i$: target vector)

example
$\mathbf{x}$: vector representation of a patient
y: has Celiac disease ($y = 1$) or not ($y = 0$)
} D consists of n "labeled" patients

$$\mathbf{x} = \begin{bmatrix} \text{concentration of IgG antibody in blood} \\ \text{has HLA-DQ2 gene} \\ \text{appetite (low, medium, high)} \end{bmatrix}$$ } stack features of a patient into a vector

goal: use training examples to train a discriminative model $P(y = 1|\mathbf{x}; \boldsymbol{\alpha})$ ← model parameters
"given the features of a patient, what is probability they have Celiac disease?" $P(y=1|\vec{x})$
"train" $\implies$ tuning model parameters $\vec{\alpha}$ to fit the training data

Three types of variables:

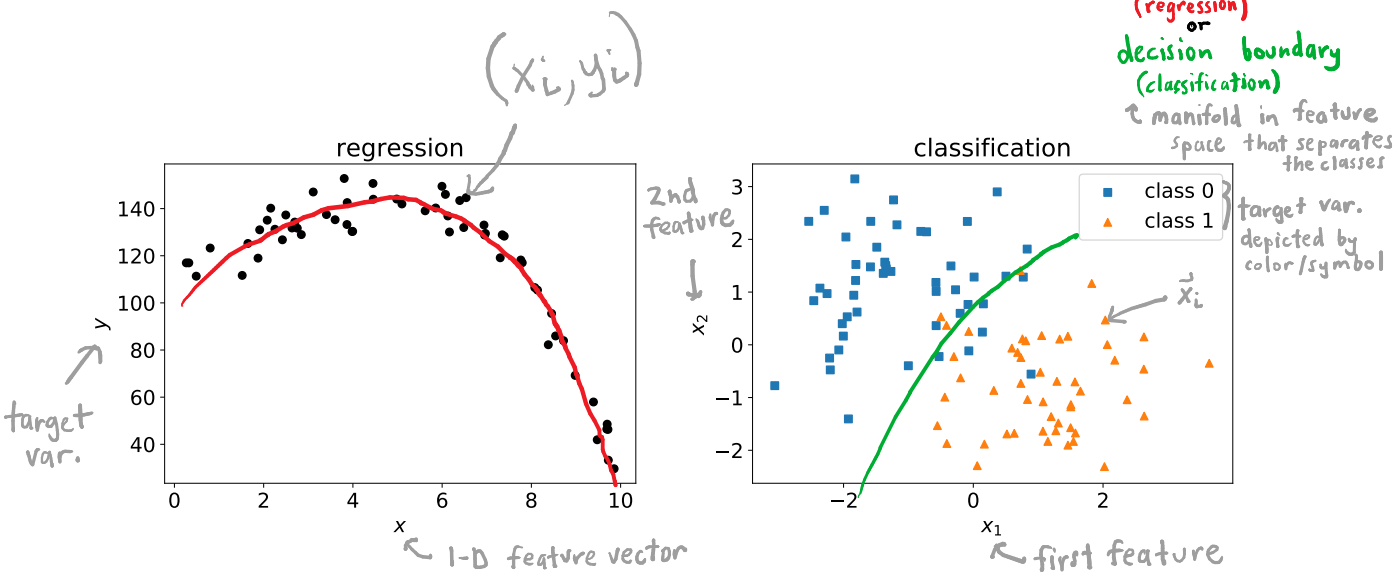
- quantitative, (approximately) continuous: quantitative variables that obey notions of order, distance
- categorical/discrete: qualitative variables (categories) without a natural order
- categorical *and* ordinal: categorical variables with a notion of order but not distance

e.g. x[1] x[2] x[3] above.

classification vs. regression

In classification, the target variables are categorical
In regression, the target variables are quantitative
↳ e.g. same $\vec{x}$ above except y := concentration of C reactive protein in blood.

conceptually, ML aims to automatically learn, from the training examples, the response surface (regression) or decision boundary (classification) ↑ manifold in feature space that separates the classes



unsupervised learning

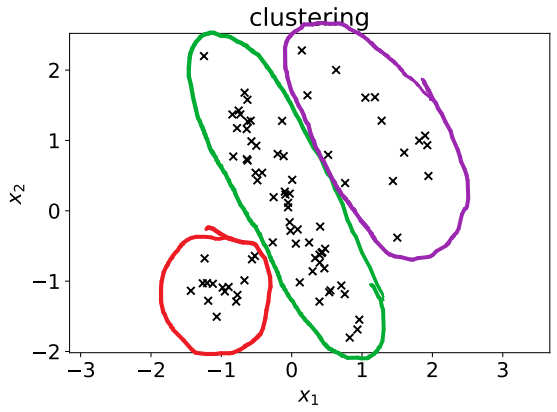
distinction: training examples are unlabeled

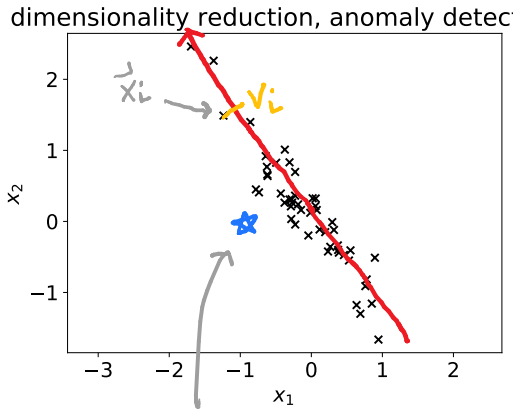
goal: find patterns/structure in data or identify groups/clusters in data (can be exploratory)

training examples: data $D := \{\mathbf{x}_1, ..., \mathbf{x}_i, ..., \mathbf{x}_n\}$ ← feature vector for example i

no target values associated with examples! (e.g. too costly to label)

- clustering : automatically group together "similar" data points
- dimensionality reduction : extract the most salient features of data, compression
- anomaly detection : if we learn the "structure" of D, assumed to consist of "normal" data, when a new data pt. comes, we can determine if it falls outside the structure of the "normal" data (i.e. if it is anomalous) e.g. credit card fraud detection





now, imagine learning the response surface, decision boundary, or "primary axis" in much higher dimensions! (can't make a plot!)

new data point. x_1, x_2 on their own not too extreme. but clearly ★ falls outside the structure of the "normal" data points (x's) ★ does not conform to normal behavior.