ECE 469/ECE 568 Machine Learning

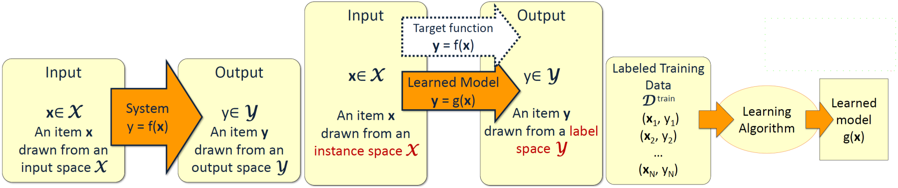
Textbook:
Machine Learning: a Probabilistic Perspective by Kevin Patrick Murphy

Southern Illinois University

August 26, 2024

Supervised Learning - Training

- Supervised learning is the form of ML most widely used in practice.
- The goal of supervised learning is to learn a mapping from inputs x to outputs y, given a labeled set of input-output pairs $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$.
- Here $\mathcal{D}$ is referred to as the training set, and N is the number of training examples.



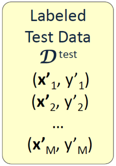
- Simply, each training input $\mathbf{x}_i$ can be a D-dimensional vector of numbers, representing, say, the height and weight of a person (features).
- $\mathbf{x}_i$ can also be a complex structured object, such as an image of person/thing, a written sentence, an email message, a time series, a molecular shape, or a graph.

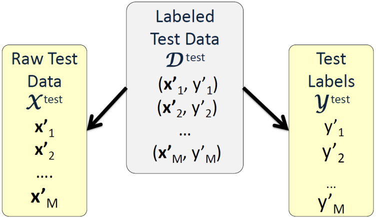
Supervised Learning - A formal definition

- We can define a training experience as a set of labeled examples (a.k.a., instances, samples) in the form: $< x_1, x_2, \cdots, x_N, y_1, \cdots, y_M >$, where x_i for $i \in \{1, \cdots, N\}$ are input variables (features/attributes), and y_j $j \in \{1, \cdots, M\}$ are the output variables.
- Our task is to learn a function $f : X_1 \times X_2, \cdots \times X_N \to Y$, which maps the input variables to the output domain.
- Our goal/objective is to minimize an error or a loss function.

Supervised Learning - Testing

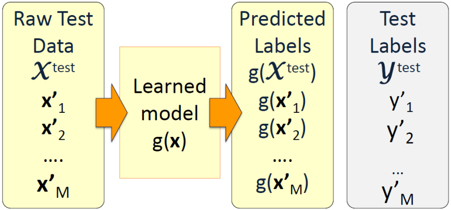
- We need to reserve some labeled data for testing.





Supervised Learning - Testing

- We can apply the trained model to the raw test data.
- Performance can be evaluated by comparing predicted labels against the test labels.



What is data?

- We can define data as a collection of examples and their features.
- Attributes and covariates are synonyms for features.

Age	Job?	City	Rating	Income
23	Yes	Van	A	22,000.00
23	Yes	Bur	BBB	21,000.00
22	No	Van	CC	0.00
25	Yes	Sur	AAA	57,000.00
19	No	Bur	BB	13,500.00
22	Yes	Van	A	20,000.00
21	Yes	Ric	A	18,000.00

"feature"

"example"

- Terminology:
  - Columns are called input variable, features or attributes.
  - Rows are called examples or samples

Types of Data

- Categorical features – > come from an unordered set:
  - Binary –> job? (yes/no)
  - Nominal –> city
- Numerical features –> come from ordered sets:
  - Discrete counts –> age
  - Ordinal –> rating
  - Continuous/real valued –> height

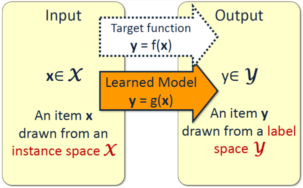
Terminology:

Input variables, Features or Attributes				
Tumor size	Texture	Perimeter	Outcome	Time
14.2	113	13.65	N	34
15.4	117	92.50	N	39
16.1	122	33.33	R	40
15.0	111	8.99	N	65

N= Non re-occurrence and R= Re-occurrence

- Firs three columns are called input variable, features or attributes.
- The last two columns are output variables (what we are trying to predict).
- The rows are called examples, samples or instances.
- Whole table is a data set.
- The problem of predicting re-occurrence/non re-occurrence is a (binary) classification problem.
- The problem of predicting time is a regression problem.

Representing data



- The kind of supervised learning task that we are dealing with is determined by the label space $\mathcal{Y}$.
- If the output labels $\mathbf{y} \in \mathcal{Y}$ are categorical, then
  - Binary classification: Two possible labels
  - Multiclass classification: K possible labels
- If output labels $\mathbf{y} \in \mathcal{Y}$ are numerical, then
  - Regression (linear/polynomial): Labels are continuous valued and task is to learn a linear/polynomial function $f(x)$
  - Ranking: Labels are ordinal, and task is to learn an ordering $f(x_1) > f(x_2)$ over input

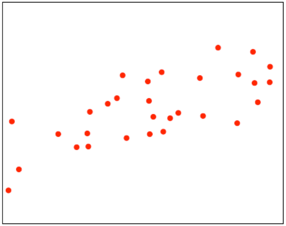
Supervised Learning - A formal definition

- Let $\mathcal{X}$ be the space of the input variables.
- Let $\mathcal{Y}$ be the space of the output variables.
- Given a data set $\mathcal{D} \in \mathcal{X} \times \mathcal{Y}$, find a mapping function $f : \mathcal{X} \rightarrow \mathcal{Y}$ such that $f(x)$ is a good predictor for the output variables y.
- Here, f is called a hypothesis.
- Supervised learning problems are categorized based on the type of the output domain.
  - If $\mathcal{Y} \in \mathbb{R}$, this is called a regression problem.
  - If $\mathcal{Y}$ is a categorical variable, it is called a classification problem
  - Generally, $\mathcal{Y}$ could be a lot more complex (graph, tree, etc). Then the underlying problem is called structured prediction.
- Typically, the parametric function f is in the hypothesis class $\mathcal{H}$.

Steps to solving a supervised learning problem

1. Decide what the input-output pairs are.
2. Decide how to encode inputs and outputs –> This defines the input space $\mathcal{X}$, and the output space $\mathcal{Y}$.
3. Choose a class of hypotheses/representations $\mathcal{H}$.
4. Choose an error function (cost function) to define the best hypothesis
5. Choose an algorithm for searching efficiently through the space of hypotheses.

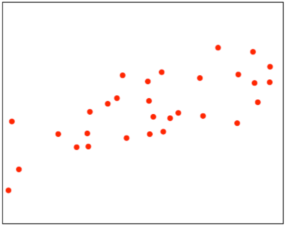
What hypothesis class should we pick?



x	y
0.86	2.49
0.09	0.83
-0.85	-0.25
0.87	3.10
-0.44	0.87
-0.43	0.02
-1.10	-0.12
0.40	1.81
-0.96	-0.83
0.17	0.43

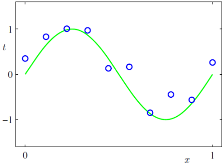
When $\mathcal{Y} = \mathbb{R}^N$, a supervised learning problem can be viewed as a regression (curve fitting) problem.

What hypothesis class should we pick?

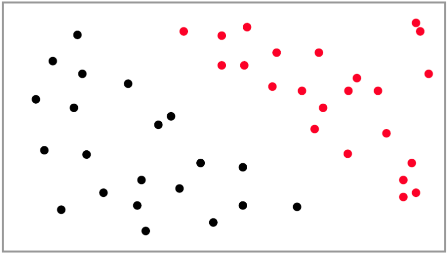


x	y
0.86	2.49
0.09	0.83
-0.85	-0.25
0.87	3.10
-0.44	0.87
-0.43	0.02
-1.10	-0.12
0.40	1.81
-0.96	-0.83
0.17	0.43

When $\mathcal{Y} = \mathbb{R}^N$, a supervised learning problem can be viewed as a regression (curve fitting) problem.



What hypothesis class should we pick?



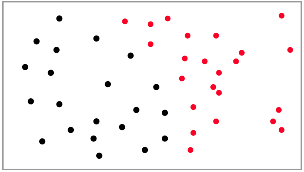
- This is linearly separable.
- Linearly separable means if f is a linear function of $\mathbf{x}$, we can perfectly fit the training data.

$$f(\mathbf{x}, \boldsymbol{w}) = \mathrm{sgn}(\boldsymbol{w}^T \mathbf{x})$$

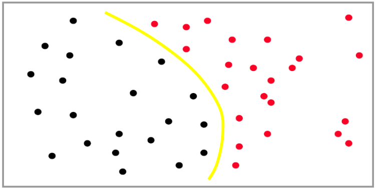
- If $\boldsymbol{w} = [w_0, w_1, w_2]^T$ and $\mathbf{x} = [1, x_1, x_2]^T$, then

$$f(\mathbf{x}, \boldsymbol{w}) = \mathrm{sgn}(w_0 + w_1 x_1 + w_2 x_2) \in \{+1, -1\}$$

What hypothesis class should we pick?



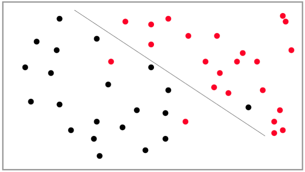
- Is this linearly separable?



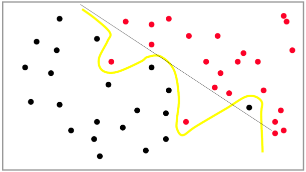
- No. This is quadratically separable.

$$f(\mathbf{x}, \boldsymbol{w}) = \operatorname{sgn}(w_0 + w_1 x_1 + w_2 x_2 + w_3 x_1^2 + w_4 x_2^2 + w_5 x_1 x_2) \in \{+1, -1\}$$

What hypothesis class should we pick?

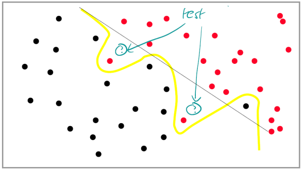


- Is this noisy/ mislabeled data?



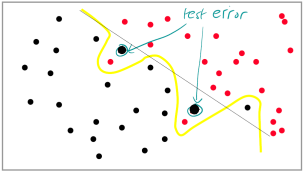
An overly flexible function memorizes irrelevant details of training set.

What hypothesis class should we pick?



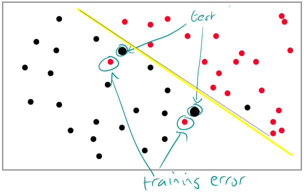
- Try to predict label of green points.
- Can overfitted functions predict test data?
- No. Overfitted functions DO NOT predict test data accurately.

What hypothesis class should we pick?



- Overfitted functions DO NOT predict test data accurately.
- Test points are mis-predicted.
- Hence, overfitting yields "TEST ERRORS".

What hypothesis class should we pick?



- Underfitted functions DO NOT predict some of the training data accurately.
- Thus, underfitting yields "TRAINING ERRORS".

What hypothesis class should we pick?

- We may trade-off "simplicity" for "accuracy of model fit".
- Moreover, if two models fit the data equally well, we may pick the simpler one.