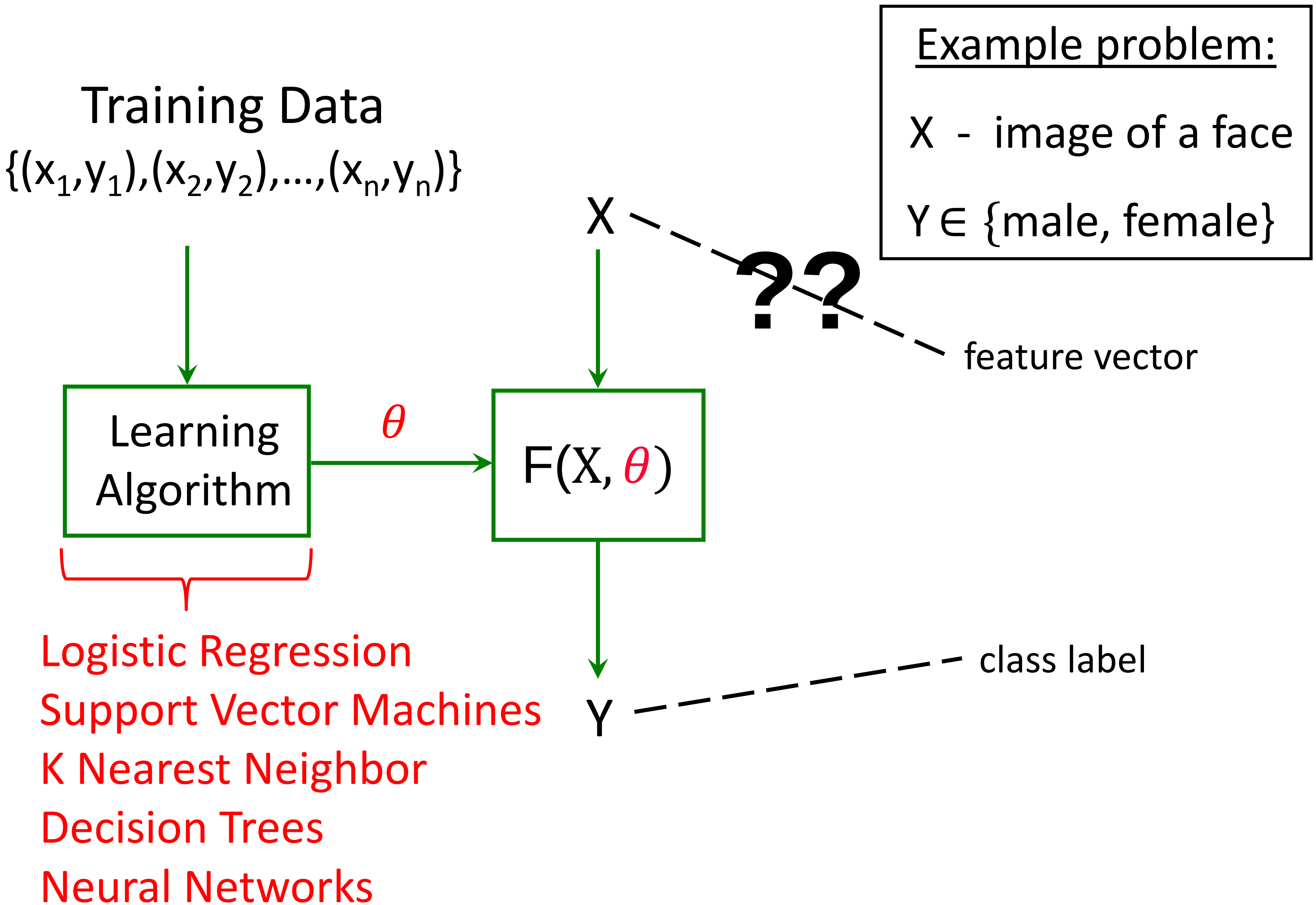
Lecture #2: Input Representation, Abstract ML Algorithm, and Supervised Learning Settings

Janardhan Rao (Jana) Doppa

School of EECS, Washington State University

Learning for Simple Outputs



Input examples: Representation

- In ML, our input examples (emails, text documents, images) are often represented as real-valued vectors: $x \in R^d$
  - each co-ordinate of x is called a feature

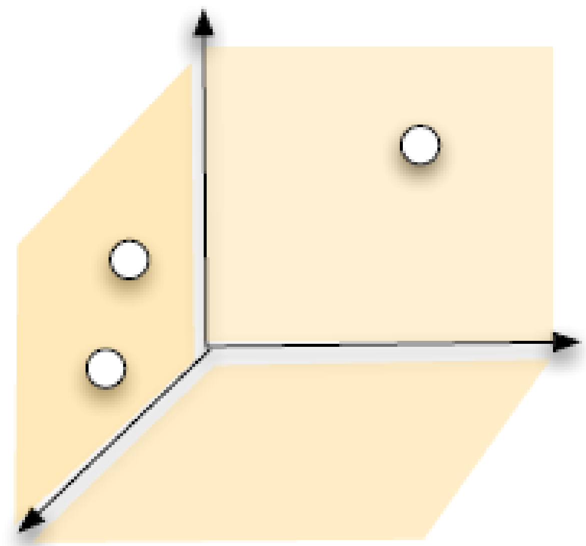
- Some examples
  - Bag-of-words representation of text
  - Histograms of colors in image
  - Sound frequency histogram

Input examples: Representation

- Bag-of-words model
  - sentences to points

1. To be, or not to be,
2. To be a woman,
3. To not be a man

To	be	or	not	woman	a	man
2	2	1	1	0	0	0
1	1	0	0	1	1	0
1	1	0	1	0	1	1

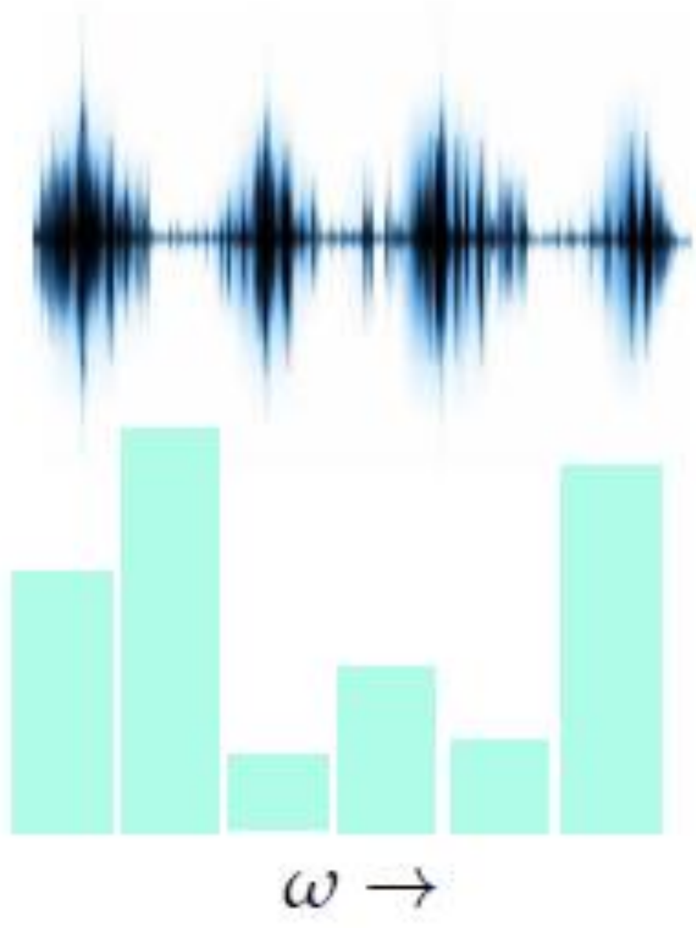


Input examples: Representation

- Histogram of colors in image

Input examples: Representation

- Sound frequency histogram



Recap of last lecture

- **Different learning paradigms**
  - Supervised, semi-supervised, unsupervised, active, and reinforcement learning
- **Representation of Input objects**
  - set of features or feature vectors
  - some examples (e.g., Bag-of-Words for text)
- **Abstract machine learning algorithm**

Overview of ML Algorithms

- There are lot of machine learning algorithms

- Every machine learning algorithm has three components
  - **Representation**
  - **Evaluation**
  - **Optimization**

Representation: Examples

- Linear hyper-planes
- Decision trees
- Sets of conjunctive / logical rules
- Graphical models (Bayes/Markov nets)
- Neural Networks
- …

Evaluation: Examples

- Accuracy
- Precision and recall
- Squared error
- Likelihood
- Cost / Utility
- Margin
- Entropy
- …

Optimization: Examples

- **Combinatorial Optimization**
  - greedy search, dynamic programming
- **Convex Optimization**
  - gradient descent, co-ordinate descent
- **Constrained Optimization**
  - linear programming, quadratic programming
- …

Generative vs. Discriminative Learning

- Training data: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ drawn from a joint distribution $\boldsymbol{P(X, Y)}$
- **Generative learning**: learn the distribution $\boldsymbol{P(X, Y)}$
  - "what controls the rise and fall of the stock prices?"
- **Discriminative learning**: learn the conditional distribution $\boldsymbol{P(Y|X)}$
  - "will there be a rise in the stock prices today evening?"

$$\boldsymbol{P(Y|X) = \frac{P(X, Y)}{P(X)}}$$

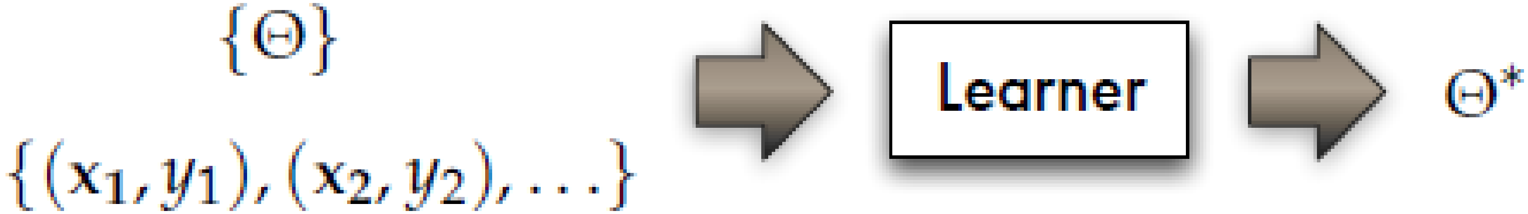
Parametric vs. Non-Parametric Learning

- **Parametric learning**
  - define a space of models parameterized by a fixed number of parameters
  - find model that best fits the data (by searching over the parameters)
  - Example: logistic regression

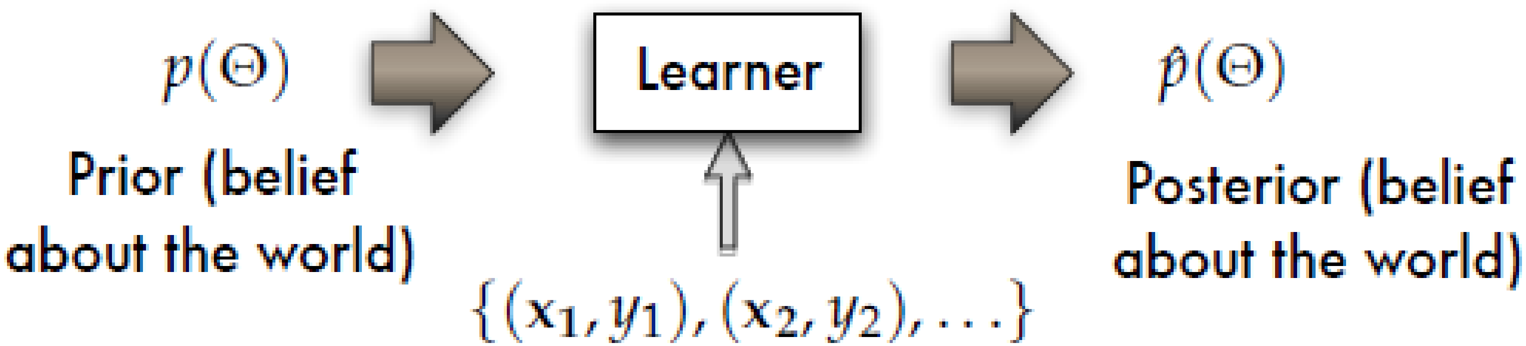
- **Non-Parametric learning**
  - define a space of models that can grow in size with data
  - find model that best fits the data
  - “Non-parametric” means “Not-fixed”
  - Example: decision trees

Non-Bayesian vs. Bayesian Learning

- **Non-Bayesian learning**



- **Bayesian learning**



Θ^* is a point estimate.

$p(\Theta)$ is a distribution over possible worlds

Machine Learned Programs: Errors

- **Approximation Error**
  - Error due to restricted hypothesis class (representation)
- **Estimation Error**
  - Error due to finite training samples
- **Optimization Error**
  - Error due to not finding a global optimum to the optimization problem