Introduction to Supervised Learning

April 11, 2016

Two problems in Machine Learning:

- Unsupervised: inferring a function to describe hidden structure. Features, but no labels.

$$\begin{array}{cccc} - & x_{11} & \cdots & x_{1n} \\ - & x_{21} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ - & x_{m1} & \cdots & x_{mn} \end{array}$$

- Examples:
  - means-based clustering, hierarchical clustering, PCA, latent variable models, etc.
  - Reinforcement learning, Q-learning, value iteration.

Two problems in Machine Learning:

- Supervised: inferring a function to describe hidden structure. Features *and* labels.

$$\begin{array}{cccc} y_1 & x_{11} & \cdots & x_{1m} \\ y_2 & x_{21} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ y_n & x_{n1} & \cdots & x_{nm} \end{array}$$

- Examples: Many!

Inferring a function?

- Suppose we have n training samples:

$$\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}.$$

Inferring a function?

- Suppose we have n training samples:

$$\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}.$$

- Suppose we have a hypothesis space, say H, of candidate functions.

- Suppose we have a score function, say $f(x, y)$.

Inferring a function?

- Suppose we have n training samples:

$$\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}.$$

- Suppose we have a hypothesis space, say H, of candidate functions.

- Suppose we have a score function, say $f(x, y)$.

- A supervised learning algorithm seeks $h : X \to Y$, $h \in H$, such that:

$$h(x) \in \operatorname*{argmax}_{y} \{f(x, y)\}.$$

Three supervised problems: Regression

- Determine the relationship between a scalar dependent variable and explanatory variables;
- Assumption: y continuous and unbounded, i.e. $y \in \mathbb{R}$;
- Examples: Linear Regression, Polynomial/basis regression, Random Forest Regression, etc.

Three supervised problems: Classification

- Determine the relationship between a categorical dependent variable and explanatory variables;
- Assumption: y is in some finite set, e.g. $y \in \{a, b, c, \dots\}$;
- Logistic regression, Probit, Ordered Logit/Probit, Multinomial Logit, Conditional Logit, Naive Bayes, GLMs, etc.

Three supervised problems: other LDVs

- Determine the relationship between a limited dependent variable and explanatory variables;
- Assumption: y is numerical, but bounded, e.g. $y \subsetneq \mathbb{R}$;
- Tobit, Poisson/NB regression, Cox PH Model, GLMs, etc.

Things to consider when solving...

- What kind of label? What kind of data?
- Bias or variance?
- How much complexity is needed? How much will the data support?
- Curse of dimenionality: too many features!
- Know the data generating process.