STK-INF4000 Notes - ML Basics

Dirk Hesse

Jan. 26th 2017

Intro

Basic idea: Some *data* collected and want to draw some *conclusions* about them. E.g.

- Spam classification.
- Search result ranking.
- Sentiment analysis.
- Predictive maintenance.

Machine learning problems fall broadly into one of two categories:

- Supervised Learning
- Unsupervised Learning

Supervised Learning

- Random variable $X = (X_1, \dots, X_p)^T$.
- Output variable Y (usually *not* a vector, but could be) *or* G (group, e.g. $0, 1$).
- *Training examples* $(x_i, y_i), i = 1, \dots, n$.
- *Task:* Given a *value* for X, make a *good* prediction for Y, denoted usually $\hat{Y}$.

Examples

- Sentiment Analysis
- Demand Prediction
- Traffic Forecasting
- Drought Forecasting.
- Increasing Farm Yields.

What Can be Learned?

- Y *must* be dependent on X.
- Dependence can be *very* complex.
  - Deep learning.

Unsupervised Learning

No Y given. Just interested in properties of X.

Examples

- Grouping Things
- Recommendation
- Outlier Detection

Basic Probability

Random Variables

- Commonly denoted X or Y.
- On continuous or discrete spaces.
- Formally
  - $X : \Omega \to \mathbb{R}$
- PMF
$$f_X(x) = \mathrm{P}[X = x] = \mathrm{P}\left[\{\omega \in \Omega : X(\omega) = x\}\right]$$
- PDF
$$\mathrm{P}[a \leq X \leq b] = \int_a^b f(x)dx$$
- CDF
$$F_X : \mathbb{R} \to [0,1], \quad F_x(x) = \mathrm{P}[X \leq x]$$
- Conditional:
$$f_{Y|X}(y|x) = \frac{f(x,y)}{f_X(x)}$$

Expectation Values

Discrete

$$\mathrm{E}[X] = \sum_x x p_X(x)$$

$$\mathrm{E}[f(X)] = \sum_x f(x)p_X(x)$$

Continuous

$$\mathrm{E}[X] = \int x p_X(x) dx$$

Conditional Probability

Study, 60% women, 40% men. High, low income.

	Female	Male	
High	9%	11%	20%
Low	46%	34%	80%
	55%	45%	

0.09, 0.11, 0.46, and 0.34 are called *joint* probabilities, while 0.55, 0.45, 0.2, and 0.8 are called *marginal* probabilities. The *conditional* probability $\mathrm{P}(High|Female)$ can be calculated using the formula

$$\mathrm{P}(x|y) = \frac{\mathrm{P}(x, y)}{\mathrm{P}(y)}$$

Here, $\mathrm{P}(High|Female) \approx 16\%$, while $\mathrm{P}(High|Male) \approx 24\%$.

Decision Theory - Supervised Learning

We have

- $X \in \mathbb{R}^p$ random input vector.
- $Y \in \mathbb{R}$ output vector.
- Looking for $f(X)$ predicting Y.

Why?

- Prediction
  - Exact form of f not too important.
  - Accuracy important.
    - How good can we do?

    * Assume $Y = f(X) + \epsilon$.
    * Our estimate $\hat{f}(X) = \hat{Y}$:

    $$\mathrm{E}(Y - \hat{Y})^2 = \mathrm{E}[f(X) + \epsilon - \hat{f}(X)]^2 = \mathrm{E}[f(X) - \hat{f}(X)]^2 + \mathrm{Var}(\epsilon).$$

- Inference
  - Which parts of X are important for predicting Y?
  - What is the relationship between Y and X?
  - What is the *functional* relationship between them? Linear?
  - **Exact form** of f *is* important.

How to get f?

We need a *loss function* $L(Y, f(X))$, most commonly squared error loss, $L(Y, F(X)) = (Y - f(X))^2$.

How to choose f?

$$\mathrm{EPE}(f) = \mathrm{E}(Y - f(x))^2 = \int (y - f(x))^2 \,\mathrm{Pr}(dx, dy)$$

Now $\mathrm{Pr}(dx, dy) = \mathrm{Pr}(Y|X)\,\mathrm{Pr}(X)$, so

$$\mathrm{EPE}(f) = \mathrm{E}_X\,\mathrm{E}_{X|Y}([Y - f(X)]^2|X)$$

and

$$f(x) = \underset{c}{\mathrm{argmin}}\ \mathrm{E}_{Y|X}\left([Y - c]^2|X = x\right)$$

This yields

$$f(x) = \mathrm{E}(Y|X = x)$$

{#eq:fEYX}

K-Nearest Neighbors

KNN implements (@eq:fEYX) in a very simple way:

$$\hat{f}(x) = \frac{1}{k} \sum_{z \in N_k(x)} z\,,$$

where $N_k(x)$ are the k closest training examples to x from a given training set.

Appendix

Formal Definitions

- Sample space Ω
- Outcome $\omega \in \Omega$
- Event $A \subseteq \Omega$
- σ-Algebra $\mathcal{A}$
- Probability distribution P
  - $\mathrm{P}[A] \geq 0 \quad \forall A$
  - $\mathrm{P}[\Omega] = 1$
  - $\mathrm{P}\left[\bigcup_i A_i\right] = \sum_i \mathrm{P}[A_i]$
- Conditional probability $\mathrm{P}[A|B] = \frac{\mathrm{P}[A \cap B]}{\mathrm{P}[B]}$, if $\mathrm{P}[B] > 0$.