Introduction to Machine Leaning I

Dihui Lai

dlai@wustl.edu

March 24, 2019

CONTENT

- Supervised Learning v.s. Unsupervised Learning
- k-Nearest Neighbors
- Naive Bayes Classifier
- Overfitting & Corss-validation

Supervised Learning v.s. Unsupervised Learning

- **Supervised Leaning**: a model/algorithm that is built on a dataset that contains both input data and desired outcome. For example, logistic regression; linear regression.
- **Unsupervised Leaning**: a model/algorithm that is built on a dataset that contains both input data but no desired outcome. For example, K-mean clustering

Distance Metrics and Geometrics of Data

The distance between two data points $\vec{x}^i$ and $\vec{x}^j$ could be measured using different metrics:

- Euclidean distance

$$d_{ij} = \sqrt{\sum_{k=1}^{n} \left(x_k^i - x_k^j\right)^2}$$

- Manhattan distance:

$$d_{ij} = \sum_{k=1}^{n} \mid (x_k^i - x_k^j) \mid$$

- Winkowski distance:

$$d_{ij} = \left(\sum_{k=1}^{n} (x_k^i - x_k^j)^q\right)^{1/q}$$

k-Nearest Neighbors

- Non-parametric classification/regression machine learning algorithm.
- Very easy to implement but can be useful.
- A decision is made based on the k closest-points in the training dataset. If k is too small ($k = 1$), the decision will be very noisy. If k is too large, neighbor includes too many data from other classes.

k-Nearest Neighbors Algorithm

Load the training and test data

Set the value of k to a reasonable value

For each point in test data:

- Calculate its distance (pick an appropriate metric like Euclidean, Manhattan etc.) to all training data points
- Sort the distances from low to high
- Choose the first k points in the training data
- Assign a class based on the majority of classes present in the chosen points (or take the weighted average for regression)

Joint Probability Distribution and Chain Rule

- The joint probability distrution of two events can be discribed as

$$P(A, B) = P(A \mid B)P(B) = P(B \mid A)P(A)$$

  If A and B are two indpendent events, then we have $P(B) = P(B \mid A)$ and $P(A) = P(A \mid B)$

- **Chain Rule**: Considering n random events X_1, X_2, X_3 ... X_n, their joint probability distribution can be described as

$$\begin{aligned}
&P(X_n, \ldots, X_1) \\
&= P(X_n|X_{n-1}, \ldots, X_1)P(X_{n-1}, \ldots, X_1) \\
&= P(X_n|X_{n-1}, \ldots, X_1)P(X_{n-1}|X_{n-2}, \ldots, X_1)P(X_{n-2}, \ldots, X_1) \\
&= \ldots
\end{aligned}$$

Bayes' Theorem

Bayes' Theorem: given two random variables X and Y, the conditional probability of X given Y is expressed as:

$$P(Y \mid X) = \frac{P(X \mid Y)P(Y)}{P(X)}$$

Useful terminologies to interpret the equations: $P(Y)$ is called prior, which is the belif in Y without any other knowledge. $P(Y \mid X)$ is the posterior taking into consideration of X. $P(X \mid Y)$ is the likelihood.

In a discrete case, the probability distribution of X can be calculated as $P(X) = \sum_i P(X \mid Y_i)P(Y_i)$

Baye's Theorem Example

Rain in California

You are planning a trip to california tomorrow. Unfortunately, the weatherman has predicted rain for tomorrow. You know in southern california, it only rains 5 days each year and there is a chance the weather man makes false predictions. You searched on line and find that when it rains, the weatherman correctly forecasts rain of 90% of the time. When it doesn't rain, he incorrectly forecasts rain 10% of the time. What is the probability that it will rain tomorrow.

Baye's Theorem Example

Solution: Denote the event that the weatherman forcast a raining day as F. The probability of rain given weatherman's forcast is

$$P(1 \mid F) = \frac{P(1)P(F \mid 1)}{P(F)}$$

.

Given that we have $P(F \mid 1) = 0.9$, $P(F \mid 0) = 0.1$, $P(1) = \frac{5}{365}$ and $P(0) = 1 - P(1) = \frac{360}{365}$. $P(F) = P(F|1)P(1) + P(F|0)P(0)$, Therefore,

$$P(1 \mid F) = \frac{P(1)P(F \mid 1)}{P(F)} = \frac{\frac{5}{365} \cdot 0.9}{0.1109} = 0.111$$

Naive Bayes Classifier

The joint probability distribution of predictor $\vec{x}$ and target variable y can be written as

$$
\begin{aligned}
& p(x_1, x_2, ...x_n, y) \\
& = p(x_1|x_2, x_3, ..., y)p(x_2, x_3, ..., y) \\
& = p(x_1|x_2, x_3, ..., y)p(x_2|x_3, x_4, ..., y)p(x_3, x_4, ..., y) \\
& = p(x_1|x_2, x_3, ..., y)p(x_2|x_3, x_4, ..., y)...p(x_{n-1}|x_n, y)p(x_n|y)p(y)
\end{aligned}
$$

Assuming features are indepedent of each other but only dpendent on the target variable, then we have

$$p(x_1, x_2, ...x_n, y) = p(y)\prod_{i=1}^{N} p(x_i|y)$$

Naive Bayes Classifier

Using Bayes' theorem, we can get the conditional probability distribution of the target variable as

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

Therefore, we have

$$p(Y|X) = \frac{p(y)\prod_{i=1}^{N} p(x_i|y)}{\sum_{y} p(y)\prod_{i=1}^{N} p(x_i|y)}$$

The denominator is constant if the features are know. $p(y)$ and $p(x_i \mid y)$ can be calculated from the data. We need to find the y that maximize the $p(Y \mid X)$, i.e.

$$\hat{y} = \underset{y}{\operatorname{argmax}}\, p(y)\prod_{i=1}^{N} p(x_i|y)$$