Machine Learning
Lecture 2

Lecture by Dr. Shai Fine
Typeset by Steven Karas

2017-10-29
Last edited 20:54:42 2017-10-29

Disclaimer These lecture notes are based on the lecture for the course Machine Learning, taught by Dr. Shai Fine at IDC Herzliyah in the fall semester of 2017/2018. Sections may be based on the lecture slides written by Dr. Shai Fine.

Agenda

- Probability and Statistics for ML
- Python Overview
- Statistical Learning

1 Probability and Statistics for ML

An experiment is a sequence of random values.

Sample Space possible outcomes of an experiment. $S = \{HH, HT, TH, TT\}$

Event a subset of possible outcomes. $A = \{HH\} \quad B = \{HT, TH\}$

Pr[E] A number assigned to an event $\Pr(A)$. Axioms:

- $\Pr(A) \geq 0$
- $\Pr(S) = 1$
- For every sequence of disjoint events $\Pr(\bigcup_i A_i) = \sum_i \Pr(A_i)$
- From (1) and (2) it follows that $\Pr(\bar{A}) \geq 1 - \Pr(A)$

Frequency Statistics $\Pr(A) = \frac{|A|}{|S|}$

Joint Probability $\Pr(A, B)$ is the probability that both events will happen.

- $\Pr(A, B) \geq 0$
- $\sum_i \sum_j \Pr(A_i, B_j) = 1$
- $\sum_i \Pr(A_i, B_j) = \Pr(B_j)$
- $\sum_j \Pr(A_i, B_j) = \Pr(A_i)$

Independence Two events A and B are independent iff $\Pr(A, B) = \Pr(A)\Pr(B)$. A set of events $\{A_i\}$ is independent iff $\Pr(\bigcap_i A_i) = \prod_i \Pr(A_i)$

Conditioning If A and B are events and $\Pr(A) > 0$, the conditional probability of B given A is $\Pr(B|A) = \frac{\Pr(A,B)}{\Pr(A)}$. For a set of events $\{B_j\}$ mutually exclusive and form a partition of the sample space, it holds that $\sum_j \Pr(B_j|A) = 1$

Conditional Independence Event A and B are conditionally independent given C in case $\Pr(A, B|C) = \Pr(A|C)\Pr(B|C)$[1]

[1] As per Shai, we will likely not use this in the course

1.1 Bayes Rule

Suppose that the events $B_1, \ldots, B_k$ are mutually exclusive and cover the sample space (i.e. one of them must occur), then for any event A:

$$\Pr(B_i|A) = \frac{\Pr(A|B_i)\Pr(B_i)}{\sum_{j=1}^{k}\Pr(B_j)\Pr(A|B_j)} = \frac{\Pr(A|B_i)\Pr(B_i)}{\Pr(A)}$$

Interpretation Consider the events B_i as possible causes (i.e. hypothesis, model), and A as an observed model outcome (i.e. evidence, data)

The probability that any of the possible causes happened given that the outcome A is observed is:

$$\textbf{Posterior} = \frac{\textbf{Likelihood} \times \textbf{Prior}}{\textbf{Evidence}}$$

Prior $\Pr(B_i)$ - prior knowledge about the probability of any of the causes to occur (before observing the evidence)

Likelihood $\Pr(A|B_i)$ - the likelihood of the evidence to occur given any of the causes

Probability $\Pr(A)$ - the probability of observing the evidence

The Bayes rule inverts likelihood relation and provides a recipe to calculate the posterior probability of any of the causes given evidence. The probability of a hypothesis/model given the data.

1.2 Descriptive Statistics

Expectation or Mean Central tendency measure $\mu = \mathrm{E}[X] = \sum_{i=0}^{N} x_i p_i$

Sample Mean Sampled mean $\bar{x} = \frac{1}{n}\sum_{i=0}^{n} x_i$

Variance Dispersion measure $\sigma^2 = \mathrm{Var}(X) = \mathrm{E}[(X-\mu)^2]$

There were many slides full of formulas that I missed. I might go back over them in the future, but it's a lot to type for marginal benefit. I might do some of it later if I need it for the homework.

1.3 Point Estimation

Choose a single value (a point) to estimate an unknown parameter value.

Maximum Likelihood Estimation X consists of independent and identically distributed (iid) samples.

Likelihood of Θ given the sample X: $I(\Theta|X) = \prod_t p(x^t|\Theta)$.
Log likelihood is $\mathcal{L}(\Theta|X) = \log I(\Theta|X) = \sum_t \log p(x^t|\Theta)$.
Maximum likelihood estimator (MLE) is $\Theta^* = \arg\max_\Theta \mathcal{L}(\Theta|X)$.

2 Data Prep

Data cleansing Filling in missing values, smoothing noisy data, indentifying/removing outliers, resolving inconsistencies.

Data transformation Normalization and aggregation

Data reduction Obtains reduced representation in volume but produces the same or even better predictive performance

Data discretization Part of data transformation but with particular importance

2.1 Missing data

We assume that all missing data is random, and that it is missing at random. There are strategies for dealing with non-random missing data, but they are out of scope.

Three standard solutions:

- Resample the missing data
- Drop instances with missing features
- Impute missing data

2.1.1 Single value imputation

Impute missing features as a single value. Mean/Median/Mode/Max/Min. May impute a single value per class.

This strategy may modify the original distribution.

2.1.2 Multivalue imputation

Create virtual instances for each possible combination of missing features. This can also be imputed by class.

2.1.3 Closest Fit; Collaborative Filtering

Impute missing features per instance based on similar rows.

$$\text{distance}(x, y) = \sum_{i=1}^{n} \text{distance}(x_i, y_i)$$

$$\text{distance}(x_i, y_i) = \begin{cases} 0 & \text{if } x_i = y_i \\ 1 & \text{if } x_i \text{ and } y_i \text{ are symbolic and } x_i \neq y_i \text{ or } x_i =? \text{ or } y_i =? \\ \frac{|x_i - y_i|}{r} & \text{if } x_i \text{ and } y_i \text{ are numbers and } x_i \neq y_i \end{cases}$$

2.1.4 Filling Gaps

Interpolation between neighboring values. Need to sort instances first, possibly by ID, timestamp, etc.

Works well for sensor data.

2.1.5 Expectation Maximization

EM is an iterative method for computing the ML parameter estimates when there is missing data.

- Replace each missing value by an estimate
- Then estimate parameters using the "complete data"
- Re-estimate each missing value using the updated parameters
- Update parameters again
- Repeat until convergence

Example: Bivariate Guassian Given some samples $x_i \in \mathbb{R}^n$ from a joint Gaussian. In some x_i, some dimensions are lost/unobserved. We know where data is missing.

Estimate $\Pr(y, z|\mu, \Sigma)$ (gaussian) from the available data.

We want to perform maximum likelihood density estimation, even though we have some missing data.

If we knew the missing values, we would maximize the log-likelihood (o = observed, m = missing):

$$\mathcal{L}_{om}(X|\vartheta) = \sum_i \log \Pr(x_i^o, x_i^m|\vartheta) \quad \vartheta = \{\mu, \Sigma\}$$

Because we don't know what the missing values X^m are, we will maximize the log-likelihood of the observed values x^o:

$$\mathcal{L}_o(X^o|\vartheta) = \log \Pr(X^o|\vartheta) = \sum_i \log \Pr(x_i^o|\vartheta) = \sum_i \log \int \Pr(x^m, x_i^o|\vartheta) dx^m$$

Here, we're stuck with the log of an integral, but we'll construct an iterative solution over $(\vartheta_1, \vartheta_2, \ldots)$

$$\mathcal{L}_o(\vartheta_n) = \log \Pr(X^o) = \int \Pr_n(X^m|X^o) \log \Pr_n(X^o) dX^m = \ldots = Q(\vartheta_n, \vartheta_n) + H(\vartheta_n)$$

$$\mathcal{L}_o(\vartheta_n) = \log \Pr_n(X^o) = \log \int \Pr_n(X^m, X^o) dX^m = \log \int \frac{\Pr_n(X^m, X^o)}{\Pr_{n-1}(X^m|X^o)} \Pr_{n-1}(X^m|X^o) dX^m$$

Now we use Jensen's inequality: $f(\mathrm{E}[X]) \geq \mathrm{E}[f(X)]$ and $f(X)$ is concave, then we get:

$$\mathcal{L}_o(\vartheta_n) \geq \int \Pr_{n-1}(X^m|X^o) \log \Pr_n(X^o, X^m) dX^m - \int \Pr_{n-1}(X^m|X^o) \log \Pr_{n-1}(X^m|X^o) dX^m$$

$$\equiv Q(\vartheta_{n-1}, \vartheta_n) + H(\vartheta_{n-1})$$

The full equations are on slides 16-18, with an example on slide 19.

3 Next week

1. ???