Bayesian learning

BENSRHIER Nabil

December 20, 2020

1) Regarding lab 5 & lab 6: See code on Github.

1 - Naïve Bayes Classifier & m-estimate:

The NBC is a technique of classification in machine learning using numerical attributes, the rule is to pick the most probable hypothesis, i.e. the assignment of a class label $\hat{y} = C_k$ for $k \in \{1, .., K\}$ as follows:

$$\hat{y} = \text{argmax}_{k \in \{1,..,K\}} p(C_k) \prod_{i=1}^{n} p(x_i|C_k)$$

The conditional probability can be estimated directly as the following frequency:

$$p(x_i|C_k) = \frac{n_{ik}}{n_k}$$

Where n_k is the total number of samples with class C_k, and n_{ik} is the number of samples with attribute x_i and class C_k.

Limitation:
The problem is that if the sample with attribute x_i and class C_k does not present in the training set, then $n_{ik} = 0$ which means that $p(x_i|C_k)$ is zero.
Thus, the whole posterior is zero since it requires multiplying all the other probabilities by that zero. Therefore, NBC cannot well predict the class of such sample.

Example: SciKitLearn digits

We implemente an NBC based on discrete values, i.e. counts of examples falling into the different classes and attribute value groups for the priors and for the conditional probabilities. And Run this on the SciKitLearn digits data :
We get a bad **accuracy** of 0.58 due to the limitation mentioned above.

classification report:

	precision	recall	f1-score	support
0	1.00	0.45	0.62	53
1	0.84	0.52	0.64	50
2	0.93	0.57	0.71	47
3	0.20	0.87	0.33	54
4	0.97	0.52	0.67	60
5	0.97	0.45	0.62	66
6	1.00	0.66	0.80	53
7	0.95	0.65	0.77	55
8	0.64	0.65	0.64	43
9	0.70	0.53	0.60	59
accuracy			0.58	540
macro avg	0.82	0.59	0.64	540
weighted avg	0.83	0.58	0.64	540

Solution : m-estimate:

The m-estimate is an approach proposed to avoid this problem as follows:

$$p(x_i|C_k) = \frac{n_{ik} + mp}{n_k + m}$$

Where p is the prior estimate of the probability we want to determine. And m is a constant equivalent to the sample size.

Improved Example:

classification report:

	precision	recall	f1-score	support
0	1.00	0.96	0.98	53
1	0.93	0.84	0.88	50
2	0.88	0.96	0.92	47
3	0.92	0.89	0.91	54
4	0.92	0.98	0.95	60
5	0.92	0.73	0.81	66
6	0.96	0.98	0.97	53
7	0.96	0.98	0.97	55
8	0.81	0.91	0.86	43
9	0.75	0.85	0.79	59
accuracy			0.90	540
macro avg	0.91	0.91	0.90	540
weighted avg	0.91	0.90	0.90	540

Intuitively, this approach will add a small number of counts to each attribute. So when we calculate the probability to observe each attribute we never get zero even if it does not exist in the training data.
Thus, this improves our accuracy from 0.58 to 0.90.

2 - NCC vs k-means vs Gaussian NBC:

NCC is the simplest Classification algorithm since it computes the centroid for each class, and given a new sample it assigns it the class of the closest centroid (based on the euclidean distance).
k-means is an algorithm that classifies data into k clusters by minimizing the least-squares distance between the instance and the centroid which is called a hard assignment.
Instead of this hard assignment, **Gaussian NBC** uses the probability of an instance to decide the possibility of its belonging to a cluster. This method does not consider clusters as spherical and it works perfectly with any non-linear distribution.

3 - k-means vs EM for GMMs
Given the algorithms is Murphy-book, It is obvious that both algorithms have two-step update process (Expectation and Maximization step). Moreover, k-means is a particular case of the EM for GMM by forcing the soft assignment to be s a hard one using:

- **argmax** function.
- fixed covariance $\Sigma = \sigma I$.
- fixed class priors.

2) Read the article: Using EM to Learn Motion Behaviors of Persons with Mobile Robots

1 - Summary :

This article shows that people's motion can be learned since they do not move randomly through an environment. Instead, they have specific trajectories related to the environment. The paper uses an approach that applies the EM-algorithm to cluster different activities into motion patterns. The input of this method is a collection of trajectories of different length related to the environment. And since the input must have a fixed shape, they choose the same length T (the maximum of trajectories) for all trajectories (by extending Ti to T) using the linear interpolation. Then the output of this algorithm is a sequence of motion patterns that represent the behaviors of people in that environment. However, the number of these patterns is unknown, so this number is determined during the learning step. Finally, this EM-method has been implemented for data recorded with laser-range finders. The first step was to determine the input trajectories by extracting positions of people in the range scans. In the second step, they evaluate their approach and prove that is able to learn motions in a domestic residence and in an office building.

2 - Article vs Lab (5 & 6):

In lab 5 we implemented multiple classifiers, among them the **Gaussian NBC**, but in lab 6 we assumed that classes are unknown and implemented the **Gaussian NBC** by using this time the EM-algorithm which solves the problem of not being able to calculate the Maximum Likelihood estimates.
In this article, we have the same approach since the motion $\theta \in \{\theta_1, .., \theta_M\}$ for each set of trajectories are unknown, and they use the same EM-method in order to predict the motions of people in the vicinity of the robot.

3 - The challenges around the actual application of the EM-algorithm in the setting described in the article:

The first challenge that I would like to mention is the number of model components which is unknown in advance, and it is determined during the training step. Therefore, the EM-algorithm leads us to two other challenges:

- Low data likelihood, i.e. a found trajectory of low likelihood, then we increase the number of components by initializing a new model component using this trajectory. Therefore, this case improves the model evaluation.
- If a motion pattern that has low utility is found, then decrease the number of components by deleting it from the model. Thus, this case reduces the model's complexity.

Hence, this mechanism avoids EM from getting stuck in a local maxima which leads to non-accurate predictions of human behavior.