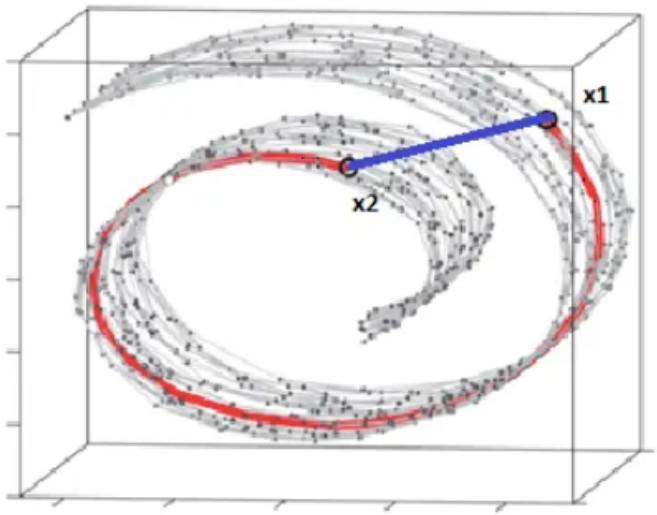
Important Issues Regarding Machine Learning

Table of Contents

Geodesic Distance

- The geodesic distance is the distance between two points following the path available/possible between them.
- The Euclidean distance is the length of a straight line between the two points.
- As the following figure shows, the data forms a spiral shape. The red curve determines the geodesic distance of the two points x_1 and x_2 while the blue line determines the Euclidean distance.



Geodesic Distance

How can we compute the geodesic distance of any two points?

1. Run KNN or radius nearest neighbors algorithm (using the Euclidean distance) to form an adjacency matrix
2. Apply a shortest-path algorithm using the adjacency matrix to calculate the geodesic distance between any two points

Algorithm

- Isomap is a manifold learning algorithm that tries to preserve the geodesic distance between samples while reducing the dimension.
  - Manifolds are curved, non-Euclidean (i.e., not $\mathbb{R}^n$) spaces that locally looks like $\mathbb{R}^n$.

1. Calculate the geodesic distance between any two points
2. Form a dissimilarity matrix whose entries are the geodesic distances

Algorithm

3. Square the dissimilarity matrix and double-center it
   - A matrix is double-centered if the transformed matrix has row mean and column mean both equal zero.Given a matrix $A \in \mathbb{R}^{n \times n}$. Form

$$A_{col} = \begin{bmatrix} mean(col1) & mean(col2) & \cdots & mean(coln) \\ mean(col1) & mean(col2) & \cdots & mean(coln) \\ \vdots & \vdots & \ddots & \vdots \\ mean(col1) & mean(col2) & \cdots & mean(coln) \end{bmatrix},$$

$$A_{row} = \begin{bmatrix} mean(row1) & mean(row1) & \cdots & mean(row1) \\ mean(row2) & mean(row2) & \cdots & mean(row2) \\ \vdots & \vdots & \ddots & \vdots \\ mean(rown) & mean(rown) & \cdots & mean(rown) \end{bmatrix}.$$

   The double-centered matrix can be computed as
   $A - A_{col} - A_{row} + mean(A)$
4. Eigen-decomposite it and choose the most dominant k eigenvectors. This is similar to what we do in PCA.

Table of Contents

- Typically, machine learning is operated in the realm of Euclidean space (i.e., $\mathbb{R}^n$). However, it is pointed out that Euclidean spaces are not fit for certain datasets like hierarchical datasets that can be described by trees.

- It is better to adopt an appropriate geometry, depending on the input data. That is, operating data in non-Euclidean spaces may give us more flexibility and require less dimensions. A manifold in the Riemannian geometry can serve as a suitable solution for us.

- An important reason we want to adopt an appropriate geometry is to preserve distances between data points after embedding them from the original space to the space where the data is represented; that is, we want to have an isometric embedding.

- If distances cannot be indeed preserved, we say distortions occur. In other words, the distortion measures the quality of the embedding by evaluating how well distances are preserved.

- For some data, there may be high distortions after embedding them into an Euclidean space; however, we may achieve low distortions if embedding them into a manifold like spherical or hyperbolic spaces. Riemannian metric, which allows to compute shortest distances on the manifold, are used.

- For example, in some NLP task, we want two words that are similar in meaning (in the semantic space) to also be similar in the embedding space (Euclidean or non-Euclidean), i.e., with a low distance (Euclidean or Riemannian). On the other hand, two words that are dissimilar in meaning should be far away in the embedding space. Hence, choosing an appropriate space is of importance.

Table of Contents

- For a machine-learning model, the MSE loss is governed by three terms, the bias, the variance and the irreducible error.
- The irreducible error is the noise term, which is due to the inherent noise in the dataset or in the problem itself and thus cannot be reduced by any model. The bias and the variance are reducible errors that we can control.
- The bias refers to the error incurred by approximating a real-world problem with a simplified model.
- The variance refers to the error from the model's sensitivity to fluctuations in the training data.

- A model has high bias because it oversimplies the relations between features and target outputs and hence fails to capture important patterns. Such phenomenon is called the under-fitting.

- As a result, we can reduce bias and capture intricate patterns by increasing the model complexity. However, this may lead to higher variance since as the model gets more complex, it also becomes more sensitive to capture noise and flucatuations.

- Actually, if a model has high variace, it not only captures the underlying patterns of data but also the noise and fluctuations too much. Such phenomenon is called the over-fitting.

- To discourage overly complex models, we can apply the regularization techniques. By adding a penalty term to the model's objective function, we can effectively restrict the complexity of the model, which in turn reduces the variability and makes the model more robust to variations in the training data.

- Our ultimate goal is to strike a good balance between under-fitting and overfitting so that the model can have low bias and low variance simultaneously.

Table of Contents

- Generalization means to predict unknown patterns in the wild while memorization means to memorize known patterns in the training data.
- If a model has low bias, it greatly captures the underlying patterns of data. However, such model may have high variance since it not only overfits the data too much but also sticks to noise and fluctuations. Such model can memorize data well but focuses on details (and noise) too much, resulting in a failure to catch the major trend and make a good generalization.

- If a model has low variance, it focuses on the major trend so that it won't be influenced by noise too much. However, such model may be too simple and is not able to learn relations between data well enough, which means it tends to under-fit the data. Such model only generalizes the major trend among data but cannot memorize information well enough, which gives rise to a high bias.

- Hence, attempting to strike a good balance between the bias and the variance amounts to managing to memorize and generalize well simultaneously. There are three possible paradigms to achieve this goal.

  1. Generalize first, memorize later
  2. Generalize and memorize simultaneously
  3. Generalize with machines and memorize with humans

Generalize first, memorize later

- The pre-training and fine-tuning approach adopted by BERT can somehow solves the generalization/memorization problem.
- After pre-training, the model starts to learn simple and generic patterns underlying the tranining data during fine-tuning. At this phase, the model learns to generalize. After simple features have been learned, the model starts to memorize specific patterns (including noise), which corresponds to the memorization phase.

Generalize first, memorize later

- Such characteristic can be verified by observing the learning curves. If the dataset is contaminated with more noise, the validation accuracy during the memorization phase will drop more steeply. Also we can deduce from the learning curves that BERT is able to memorize a specific training example only once it has seen that example a certain number of times.

- As a whole, BERT generalizes first and memorizes later. Furthermore, it is claimed that this phenomenon is due to the implementation of pre-training. If we simply randomly initialize a BERT model, it does not exhibit these training phases.

Generalize and memorize simultaneously

- In modern recommendation systems, it is important to have both memorization and generalization. Without memorization, the customers will be frustrated with unlike items. Without generalization, the customers will be bored with a bundle of similar items.

- In 2016, Heng-Tze Cheng and collaborators proposed "wide and deep learning". The key idea is to build a single neural network that has both a deep component (a deep neural net with dense embedding features as inputs) for generalization as well as a wide component (a linear model with a large number of cross-categorical features as inputs) for memorization.

Generalize with machines and memorize with humans

- In 2014, Chong Sun and collaborators from Walmart labs proposed a production system for large-scale e-commerce item classification. Such system incorporates human decisions to handle a large number of edge cases with little training data. Human analysts can write rules to cover these cases precisely. Then these rules can later be merged and fused into the model training so that after some time the model can learn these new specific patterns.

Table of Contents

Reference


Dimension Reduction using Isomap
https://medium.com/data-science-in-your-pocket/dimension-reduction-using-isomap-72ead0411dec


Machine Learning in A Non-Euclidean Space
https://towardsdatascience.com/machine-learning-in-a-non-euclidean-space-99b0a776e92e


Striking the Balance: Bias and Variance in Machine Learning and the Golf Analogy.
https://vocal.media/education/striking-the-balance-bias-and-variance-in-machine-learning-and-the-golf-analogy


Machines That Learn Like Us: Solving the Generalization-Memorization Dilemma
https://towardsdatascience.com/solving-machine-learnings-generalization-memorization-dilemma-3-promising-paradigms-ab9c236add3e