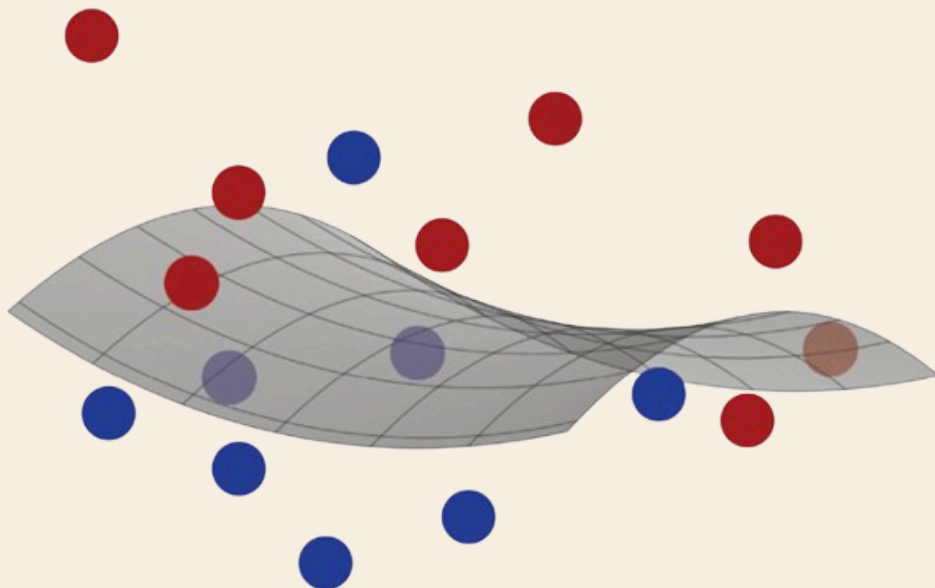


Foundations of Machine Learning

DAY - 5


Learning Stages - Example





Learning Stages - Example

Created By: **Birva Dave**

 medium.com/@birva1809

 github.com/Birva1809

 linkedin.com/in/birva-dave

 birvadave1809@gmail.com

Let us now walk through the entire learning pipeline for our spam detection example.

1. Split the Labeled Data

- We start with a dataset of emails, each labeled as spam or not spam.
- First step: randomly split this dataset into three parts:
 - Training Sample – used to train the model
 - Validation Sample – used to tune hyperparameters
 - Test Sample – used to evaluate final performance
- How much data we assign to each part depends on things like:
 - The number of hyperparameters we need to tune (more = bigger validation set)
 - Total size of the dataset
 - If the dataset is small, we usually give more data to training, because model performance depends heavily on training quality.

2. Choose and Extract Features

- For each email, we define and extract a set of relevant features (e.g., word frequency, subject line patterns, etc).
- This step is crucial — the learning algorithm only sees the features, not the raw email.
 - Good features make it easier for the algorithm to find patterns.
 - Bad or noisy features confuse the model and reduce performance.


⚠ Feature selection is up to us (the user), and our prior knowledge about spam patterns plays a big role here.


This can dramatically impact how well the final model performs.


Learning Stages - Example

Created By: **Birva Dave**

 medium.com/@birva1809

 github.com/Birva1809

 linkedin.com/in/birva-dave

 birvadave1809@gmail.com

3. Train the Algorithm

- Now we run the learning algorithm on the training data using the selected features.
- The algorithm has some free parameters (aka hyperparameters, like learning rate, model complexity, etc).
- For every combination of hyperparameters, the algorithm:
 - Picks a different hypothesis from the hypothesis set
 - Trains a model
 - Evaluates its performance on the validation sample
- We select the combination of hyperparameters that performs best on the validation set — this gives us the final hypothesis, let's call it h_0 .

4. Final Evaluation on Test Data

- Now, we freeze the model (h_0) and test it on the test sample — this is data the model has never seen before.
- We predict the labels and compare them to the actual labels using a loss function (like zero-one loss in classification).
- This gives us the test error, which tells us how well the model will perform in the real world.

Important:

We never use training or validation accuracy to report final performance.

Only the test error matters for evaluating how good the model truly is.