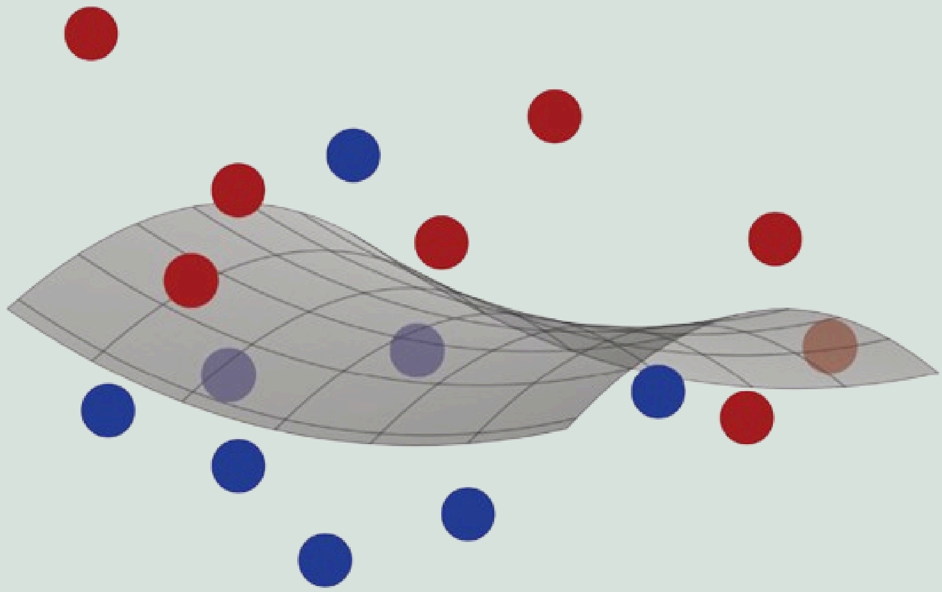
Foundations of Machine Learning

DAY - 11

Generalities



Generalities

Created By: **Birva Dave**
medium.com/@birva1809
github.com/Birva1809
linkedin.com/in/birva-dave
birvadave1809@gmail.com

In supervised learning, the core objective is to learn a hypothesis from a labeled dataset such that the hypothesis generalizes well to unseen data. There are two primary scenarios to consider in this context: deterministic and stochastic.

Deterministic vs Stochastic Scenarios

- In the general supervised learning setup, the data distribution D is defined over X × Y, and the training data is an i.i.d. sample from D:
$$S = ((x_1, y_1), \ldots, (x_m, y_m))$$
- The goal is to find a hypothesis $h \in H$ that minimizes the generalization error, defined as:
$$R(h) = P_{(x,y)} \quad D\, [h(x) \neq y] = E_{(x,y)} \quad D\, [1\_\{h(x) \neq y\}]$$
- In this setting, the output label is not deterministically assigned by the input; rather, it is drawn from a conditional distribution P[y | x]. This probabilistic labeling models many real-world tasks. For example, predicting gender from height and weight involves ambiguity—several input combinations may correspond to multiple likely labels.
- The deterministic scenario, on the other hand, assumes there exists a target function f: X → Y such that:
$$y_i = f(x_i) \text{ for all } i \in \{1,\ldots,m\}$$
- Here, the distribution D is only over the input space X, and the label of each input is uniquely determined. While this setup simplifies theoretical analysis, it may not always reflect the uncertainty in real-world data.

Generalities

Created By: **Birva Dave**
medium.com/@birva1809
github.com/Birva1809
linkedin.com/in/birva-dave
birvadave1809@gmail.com

Agnostic PAC Learning

- The agnostic PAC (Probably Approximately Correct) learning framework generalizes the classical PAC model to handle the stochastic case where no perfect hypothesis may exist within the hypothesis class H.
- A learning algorithm A is said to be an agnostic PAC learner if, for any distribution D over X×Y, and for any $\epsilon>0$ and $\delta>0$, it produces a hypothesis $h_S \in H$ such that:

$$\mathbb{P}_{S\sim\mathcal{D}^m}\left[R(h_S) \leq \min_{h\in\mathcal{H}} R(h) + \epsilon\right] \geq 1 - \delta$$

- Here, h_S is the hypothesis learned from sample S, and the guarantee is that its error is close to the best possible error within H, with high probability. If the algorithm also runs in time polynomial in $1/\epsilon$, $1/\delta$,, and the size of the input, it is considered efficient.

Bayes Error and Noise

- In deterministic settings, the existence of a perfect target function implies that a hypothesis with zero error is achievable:

$$\exists f \in \mathcal{H} \text{ such that } R(f) = 0$$

- However, in the stochastic setting, no hypothesis can achieve zero error due to the intrinsic uncertainty in labels. The Bayes error represents the lowest possible error achievable by any measurable function:

$$R^* = \inf_{h \text{ measurable}} R(h)$$

Created By: **Birva Dave**
medium.com/@birva1809
github.com/Birva1809
linkedin.com/in/birva-dave
birvadave1809@gmail.com

Generalities

- A hypothesis that achieves this minimum is called a Bayes classifier, defined pointwise as:

$$h_{\text{Bayes}}(x) = \arg \max_{y \in \{0,1\}} \mathbb{P}[y|x]$$

- This classifier predicts the label with the highest conditional probability. The associated error at point x is:

$$\min\{\mathbb{P}[y = 0|x], \mathbb{P}[y = 1|x]\}$$

- Taking the expectation over all x D, we obtain the Bayes error:

$$R^* = \mathbb{E}_{x \sim \mathcal{D}} \left[\min\{\mathbb{P}[0|x], \mathbb{P}[1|x]\}\right]$$

- This error reflects the inherent noise in the task. The noise at a point x is defined as:

$$\text{noise}(x) = \min\{\mathbb{P}[y = 0|x], \mathbb{P}[y = 1|x]\}$$

- And the average noise over the distribution D equals the Bayes error:

$$\text{Noise} = \mathbb{E}_x[\text{noise}(x)] = R^*$$

Inputs where noise(x) is close to 1/2 are considered highly ambiguous or “noisy”, and present a challenge for accurate learning.

Generalities

Created By: **Birva Dave**
medium.com/@birva1809
github.com/Birva1809
linkedin.com/in/birva-dave
birvadave1809@gmail.com

Summary

- The stochastic scenario generalizes learning to include probabilistic labels, better reflecting real-world conditions.
- The agnostic PAC-learning model provides a robust theoretical framework for learning in the presence of label noise.
- Bayes error defines the theoretical lower bound on the classification error.
- Noise quantifies the ambiguity in labeling and serves as a measure of problem difficulty.

These concepts deepen our understanding of the limitations and capabilities of learning algorithms in complex, uncertain environments