Study Of Active Learning Algorithms

Kanupriya Jain
Institut Polytechnique de Paris

Supervisor:
Prof. Christophe Denis
LPSM, Sorbonne Université

Table of Contents

Definition

It is a supervised learning algorithm that categorizes the new observations into one of two classes. We will particularly focus on 1-0 classification where the output is either 1 or 0.

Examples :

- **Spam Detection:** Classifying emails as spam (1) or not spam (0).
- **Medical Diagnosis:** Determining if a patient has a certain disease (1) or not (0).
- **Sentiment Analysis:** Assessing whether a review is positive (1) or negative (0).

Types Of classifier Algorithms

1 Binary Classification

Linear Classifiers:

- Logistic Regression
- Support Vector Machines (having *kernel*="linear")
- Stochastic Gradient Descent(SGD) Classifier

Non-Linear Classifiers :

- K-Nearest Neighbours
- Decision Tree Classification
- Random Forest

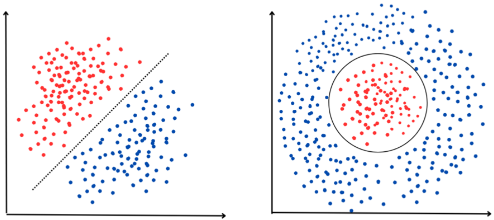


Table of Contents

Supervised Classification Setting (1/3)

- **Observation :** (X, Y) where $X \in \mathcal{X} = \mathbb{R}^d$ (Instance Space) and $Y \in \mathcal{Y} = \{0, 1\}$
- **Classifier :** $f : \mathbb{R}^d \to \{0, 1\}$
- **Risk :** $R(f) = \mathbb{P}(f(X) \neq Y)$
- **Bayes Classifier :** $f^* \in \arg\min_f R(f)$
- **Bayes Risk :** $\mathcal{R}^* = \inf_f R(f)$
- $\eta(x) = \mathbb{P}(Y = 1|X = x)$
- The Bayes classifier and Bayes risk in our case are given by

$$f^*(x) = 1_{\{\eta(x) \geq \frac{1}{2}\}} \text{ and } \mathcal{R}^* = \mathbb{E}[min\{\eta(X), 1 - \eta(X)\}]$$

.

Supervised Classification Setting (2/3)

1. **Data:** i.i.d. sample $(X_1, Y_1), \dots, (X_n, Y_n)$ from an unknown probability distribution.
2. **Objective :** To find a map from *Instance Space* $\mathcal{X}$ to set of outputs $\mathcal{Y} = \{0, 1\}$ called the *label space.*

Notations:

- $\mathcal{G}_\Theta = \{f_\theta : \theta \in \Theta\}$, set of classifiers of the form $f_\theta = 1_{\{\eta_\theta(x) \geq \frac{1}{2}\}}$
- The loss function that we will consider is $\ell(y, z) = 1_{\{y \neq z\}}$
- We can observe that

$$\mathcal{R}^* = \mathcal{R}(f^*) = 1 - \mathbb{E}[g^*(X)]$$

where $g^*(.) = max\{\eta(.), 1 - \eta(.)\}$ is called the *score function.*

Supervised Classification Setting (3/3)

- Parametric class of regression functions $\mathcal{F} = \{\eta_\theta : \theta \in \Theta\}$
- Emperical Risk Minimizer is defined as

$$\hat{\theta}_n \in \arg\min_{\theta\in\Theta} \hat{\mathcal{R}}_n(f_\theta) \quad \text{where } \hat{\mathcal{R}}_n(f_\theta) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{f_\theta(x_i)\neq y_i\}}$$

and its oracle counterpart is defined as

$$\theta^* \in \arg\min_{\theta\in\Theta} \mathcal{R}(f_\theta) \text{ where } \mathcal{R}(f_\theta) = \mathbb{E}[1_{\{f_\theta(\mathcal{X})\neq\mathcal{Y}\}}]$$

Table of Contents

Emperical Risk Minimization (1/2)

Let $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ be *i.i.d.* data points.

Assumptions :

- Θ, the class of parameters admits an ϵ-cover Θ_ϵ.
- We assume that the Bayes predictor $f^* = f_{\theta^*} \in \mathcal{G}_\Theta$.
- We also assume that the $\hat{\theta} \in \Theta_\epsilon$
- **Lipschitz Condition** For any x, $||\eta_\theta(x) - \eta_{\theta'}(x)|| \leq ||\theta - \theta'||\ ||x||$
- **Compactness** $\mathcal{X} \subset \overline{B}(0, R)$

Emperical Risk Minimization(2/2)

Theorem

For the above supervised learning setting, assumptions and ERM the following inequality is satisfied

$$\mathbb{E}[R(f_{\hat{\theta}}) - R(f^*)] \leq O\left(\sqrt{\frac{\log(2|\Theta_\epsilon|)}{n}} + \epsilon\right)$$

Remark:

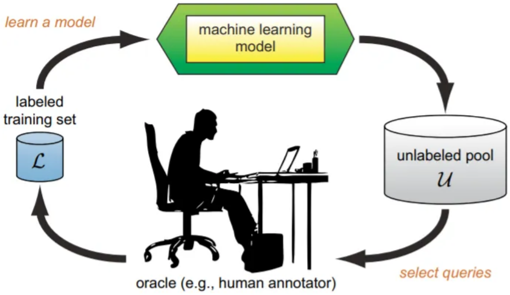
- If $\Theta \subset \mathbb{R}^M$ and is compact, then $|\Theta_\epsilon| \leq c\left(\frac{1}{\epsilon}\right)^M$
- The case discussed above is not used in practice due to our choice of non-convex loss function.

Table of Contents

Definition

Active Learning aims to reduce the number of labeled data required for training and selecting the data which needs to be labeled.

Active Learning Framework :



Motivation and Strategies

Motivation :

- Acquiring labeled data can be costly and time-consuming.
- In many domains, such as medical imaging or legal document analysis, expert labeling is required, increasing costs.

Sampling Strategies:

1. **Stream based Sampling:** An active learning technique for training models on continuous data streams, where each sample is sequentially evaluated for labeling.
2. **Pool based Sampling:** It begins with a large pool of unlabeled data and then ranks all samples in the pool and then selects the best ones to query.

Disagreement Region

One of the key principles of active learning is to identify at each step the region of the instance space where the label requests should be made, called *Uncertain Region*, also known as *Disagreement Region*

Ways to find Uncertain Region :

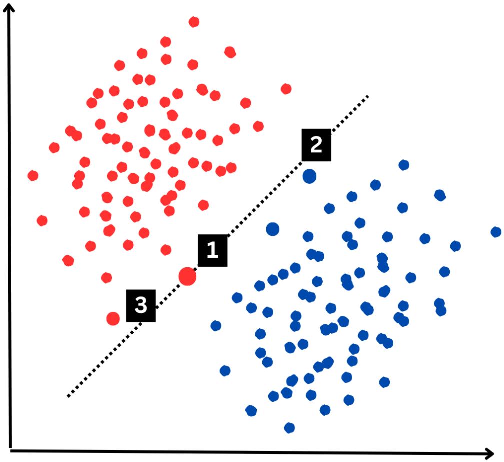
1. Uncertainty Sampling
2. Query By Committee (QBC)
3. Rejection

Uncertainty Sampling

- This algorithm involves identifying and ranking unlabeled data points where the model's predictions are least confident, calculated by $|\eta(x) - 0.5|$.
- The instances with the smallest margins are the most informative and are selected for querying

Uncertainty Sampling

Kanupriya Jain



Query Strategies:

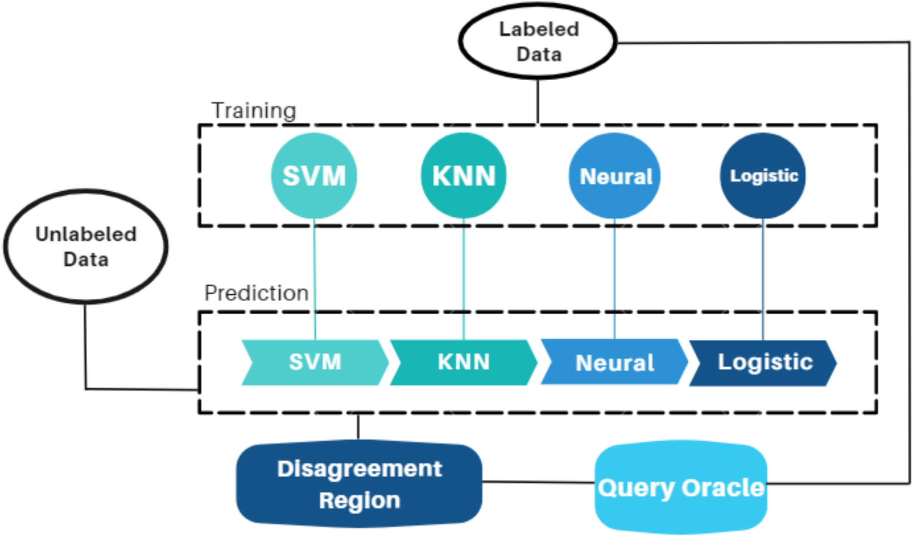
4 Active Learning

Query By Committee (QBC)

- In this approach, a committee of models is trained on the current labeled dataset, each representing different hypotheses.
- The models vote on the labeling of query candidates (unlabeled examples). Instances where the committee disagrees are considered the most informative.

Query By Committee (QBC)

4 Active Learning



Rejection

- In this learning method instead of simply assigning a class label to every instance it encounters, the model may choose to "reject" predicting a label for instances where its confidence falls below a certain threshold.
- It has valuable in applications like medical diagnosis, where an incorrect prediction can have serious consequences.

Table of Contents

Active Learning Strategy in a nutshell : budget N

- $M_0 \geq 1$, training initial estimator $\hat{\eta}_0$ on $B = M_0$ points and $A_0 = \mathcal{X}$
- while $B \leq N$
    1. $l \leftarrow l + 1$ Build new $A_l \subset A_{l-1}$
    2. Sample (X_i, Y_i), $i = 1, \dots, M_l$ $s.t$ $X_i \sim \pi(\cdot|A_l)$ and build $\hat{\eta}_l$
    3. Update $\hat{\eta} = \sum_{j=0}^{l-1} \hat{\eta}_j 1_{\{A_j \setminus A_{j+1}\}} + \hat{\eta}_l 1_{A_l}$
    4. $B \leftarrow B + M_l$

Notations

- We take our initial uncertain region as $A_0 = \mathcal{X} = [0,1]^d$. We will use the score function as $g(x) = max\{\eta(x), 1 - \eta(x)\}$
- We also take fixed number of label requests N (called the budget).
- We define a sequence of positive numbers $(\epsilon_k)_{k\geq 0}$ as a *sequence of rejection rates* and another sequence $(N_k)_{k\geq 0}$ defined as $N_0 = \lfloor \sqrt{N} \rfloor$ and $N_{k+1} = \lfloor c_N N_k \rfloor$ with $c_N > 1$.
- We will construct a sequence of uncertain regions $(A_k)_{k\geq 1}$ and estimators $\hat{\eta}_k$ on A_k.

ALGORITHM

INITIALIZATION PHASE

Sample i.i.d N0 points according to Π0 = Π

DN0 = {(X1, Y1), . . . , (XN0 , YN0)}

Estimator : ˆη0'
Classifer : ‘f0’
Score function:’g0'

ITERATION PHASE

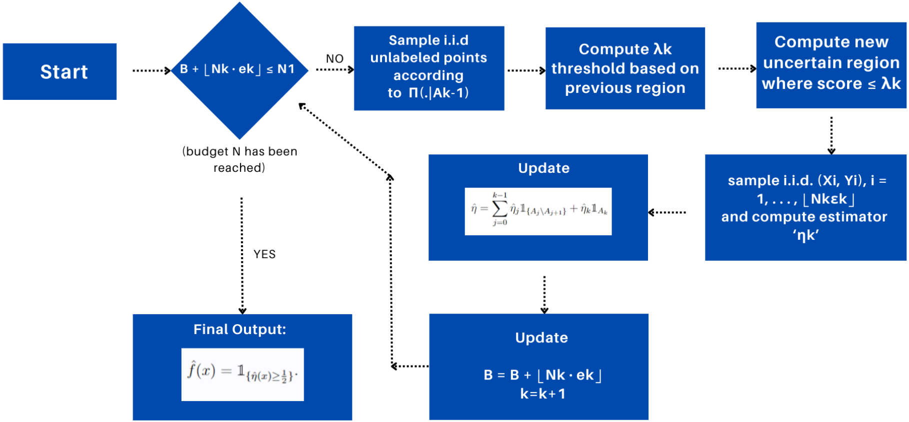
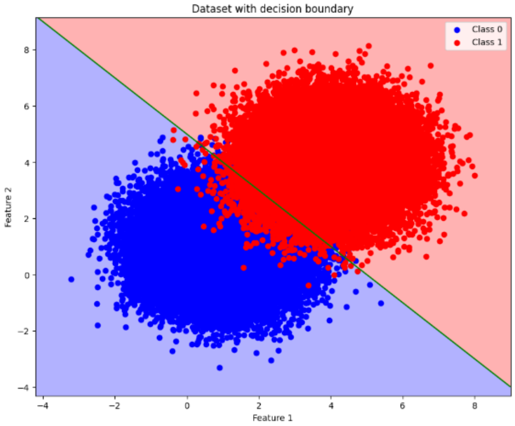


Table of Contents

Synthetic Dataset

6 Numerical Experiments



Accuracy

Kanupriya Jain

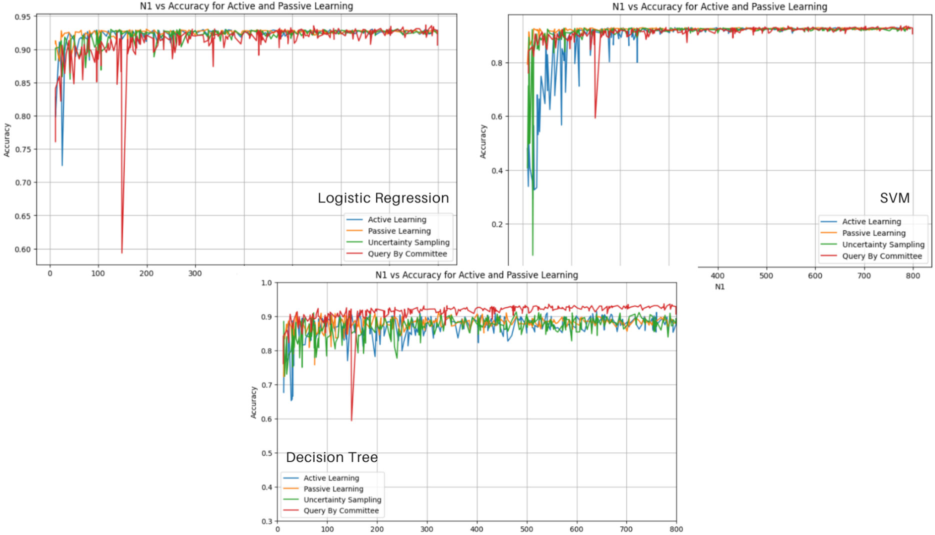
Classifier	N	Active Framework	Passive Framework	Uncertainty Sampling
Logistic Regression	400	0.9186 ± 0.0026	0.9187 ± 0.0027	0.9165 ± 0.0019
SVC	400	0.9165 ± 0.0043	0.9185 ± 0.0023	0.9169 ± 0.0034
Decision Tree	400	0.8734 ± 0.0222	0.8740 ± 0.0111	0.8781 ± 0.0202

Table: Comparison of different classifiers and frameworks.

Accuracy for QBC : 0.9158 ± 0.0046

Accuracy Vs Budget

Kanupriya Jain



Accuracy for Non-Synthetic Dataset(Stroke Prediction)

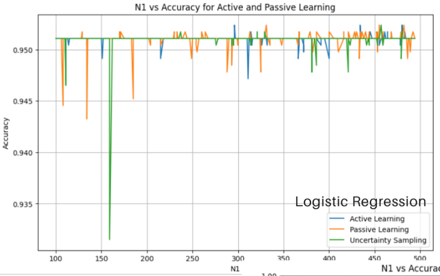
Classifier	N	Active Framework	Passive Framework	Uncertainty Sampling
Logistic Regression	500	0.9508 ± 0.0012	0.9511 ± 0.0004	$0.9511 \pm$ 2.220e-16
SVC	500	0.9509 ± 0.0008	$0.9511 \pm$ 2.220e-16	$0.9511 \pm$ 2.220e-16
Decision Tree	500	$\mathbf{0.9152 \pm 0.0141}$	0.9056 ± 0.0097	$0.9511 \pm$ 2.220e-16

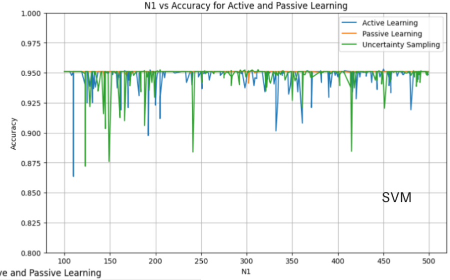
Table: Comparison of different classifiers and frameworks.

Accuracy for QBC : $0.9511 \pm$ 2.220e-16

Accuracy Vs Budget

6 Numerical Experiments





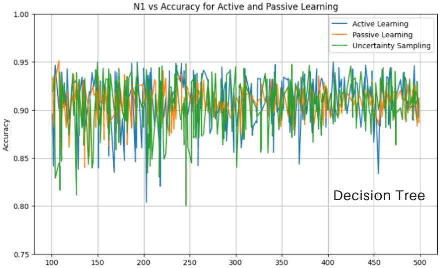


Table of Contents

Contribution

1. **Theoretical :**
    - Read the research paper *Active learning algorithm through the lens of rejection arguments* by Christophe Denis, Mohamed Hebiri, Boris Njike, Xavier Siebert (2022).
    - Proof of the theorem for Empirical Risk Minimization for passive learning framework.
2. **Practical :**
    - Simulated synthetic datasets and studied the results on passive learning framework and compared them with Bayes classifier
    - Implementation Of active learning algorithm using rejection method, Query by Committee (QBC), and Uncertainty Sampling from scratch.
    - Implementation of the above algorithms on Synthetic and Real Dataset (Stroke Prediction).
    - Comparison of the active learning results with passive learning algorithms.

Perspective

1. **Theoretical:**
   - Understand the theoretical part of the paper concerning the bound on the emperical risk.
   - Propose a similar result but in the case of parametric active learning setting
2. **Practical:**
   - Find new scenario where active learning framework is relevant.
   - Comparison of the results of active learning algorithms implemented with already existing active learning libraries online.

Table of Contents

References (1/2)

8 References

- Christophe Denis, Mohamed Hebiri, Boris Njike, Xavier Siebert (2022) - *Active learning algorithm through the lens of rejection arguments*
- A Probabilistic Theory of Pattern Recognition (1996) by *Luc Devroye, Laszlo Gyorfi, Gabor Lugosi*
- A quick overview of Active Learning
- Active Learning: Curious AI Algorithms
- Active Learning in Classification — Query Strategies
- **Dataset :** Stroke Prediction Dataset

Reference (2/2)

- Gabriel Stoltz (April 15, 2024) An *Introduction to Machine Learning [Lecture Notes]*, Institut Polytechnique de Paris
- Freund, Y., Seung, H. S., Shamir, E., Tishby, N. (1997). *Selective sampling using the query by committee algorithm*. Machine Learning, 28, 133–168

The End

Thank you for listening !