BITS-F459: Computer Vision

06 Machine Learning Basics

Dr. Kamlesh Tiwari
Associate Professor, Department of CSIS,
BITS Pilani, Pilani Campus, Rajasthan-333031 INDIA

Jan 20, 2025 ON-CAMPUS NAB 6108 @ BITS-Pilani [Jan-May 2025]

http://ktiwari.in/cv

Introduction

ML depends upon **Pattern Recognition** which corresponds to finding regularities in the data.

- There should be a pattern.
- No issues if we are unable to describe it mathematically.
- Sufficient number of examples or data is required.

Consider e-mail filtering SPAM/Not-SPAM

Assumption is that there are some words whose frequency is correlated to this filtering.

Netflix Prize (2009)

Open competition to predict user ratings for films.
Prize of USD 1 million was given to the BellKor's Pragmatic Chaos team which improved previous prediction by by ∼10.06%
(used matrix factorization)

Building Blocks

- Input: x
- Output: y
- Training data: $(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), ..., (x^{(m)}, y^{(m)})$
- $x^{(i)}$ could be a multivariate say $x^{(i)} = (x_1^{(i)}, x_2^{(i)}, ..., x_n^{(i)})$
- Concept, target function: true function

$$f : x \to y$$

- Hypothesis:

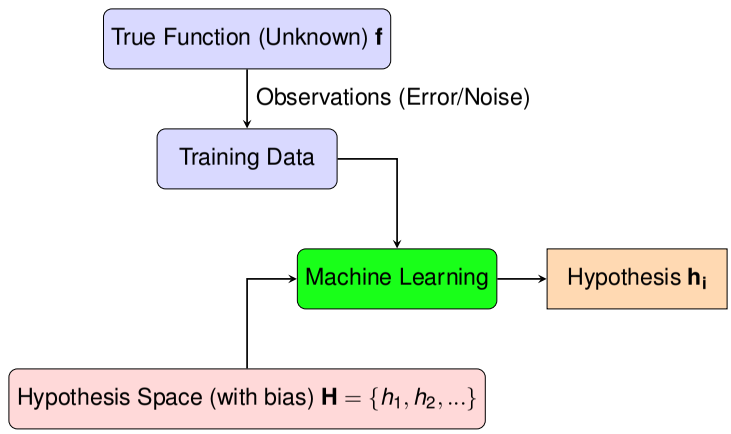
$$h : x \to y$$

- Accuracy: agreement b/w f and h

Issue is

The true function is NOT known.

The Flow of ML



A Toy model (linear)

- **The Problem:** credit approval.
- Input: $x = (x_1, x_2, ..., x_n)$
- Let x_1=accountBal, x_2=Salary, x_3=age ...
- What weights we should give w_1=0.6, w_2=0.3, w_3=-0.1 ...
- The Model

$$\sum_{i=1}^{n} w_i \times x_i = \begin{cases} > Threshold & \text{Then APPROVE} \\ otherwise & \text{DENY/REJECT} \end{cases}$$

- Simplified:

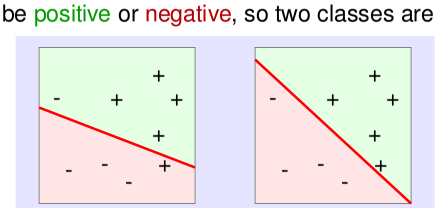
$$h(x) = sign(\sum_{i=1}^{n} w_i \times x_i - Threshold)$$

- Add an extra term x_0 (that is always 1), then

$$h(x) = sign(\sum_{i=0}^{n} w_i \times x_i) = sign(w^T x)$$

A Toy model (Contd..)

- Can you recognize $h(x) = sign(\sum_{i=0}^{n} w_i \times x_i)$
- It is a linear equation (in two dimension) or hyper plane
- Sign could be positive or negative, so two classes are **+1** and **−1**



- Vector $w = (w_0, w_2, ..., w_n)$ would be normal to the plane of linear decision boundary. (why? because dot product is $\cos\theta$)
- What could change this plane? w_i's

Learning: Use misclassified examples $x^{(k)}$ to update $w_i = w_i + \alpha y_k x_i^{(k)}$

A Toy model (What yw^Tx tells)

- Classification of a point x can be obtained from w^Tx.
  If $\boxed{w^Tx}$ is positive then x is positive, otherwise negative.
- ML assumes that data point is never on hyperplane so $w^Tx \neq 0$
- There could be two cases
  1. **CORRECT classification by the model**:
     For $y = +1$ we have $w^Tx > 0$
     For $y = -1$ we have $w^Tx < 0$
     In both the cases $yw^Tx > 0$
  2. **WRONG classification by the model**:
     For $y = +1$ we have $w^Tx < 0$
     For $y = -1$ we have $w^Tx > 0$
     In both the cases $yw^Tx < 0$

So we have a simple test

$$yw^Tx \begin{cases} > 0 & \text{Classification is correct} \\ < 0 & \text{Classification is wrong} \end{cases}$$

Hyperplane and Normal Vector

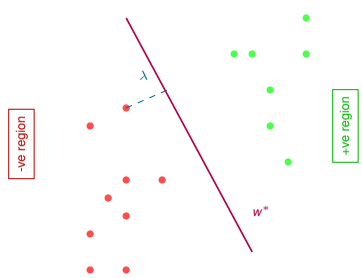
Consider hyperplane: $w_1x_1 + w_2x_2 + ... + w_mx_m = 0$

- It passes through origin O: as $(0, 0, 0, ..., 0)$ satisfies the equation
- What is $W = (w_1, w_2, ..., w_m)$? vector/point; so does $kW = (kw_1, kw_2, ..., kw_m)$
- Does $(w_1, w_2, ..., w_m)$ lies on hyperplane?
  NO. As the equation $w_1w_1 + w_2w_2 + ... + w_mw_m$ is non-zero
- Consider a point $P = (p_1, p_2, ..., p_m)$ that lies on hyperplane. Then $w_1p_1 + w_2p_2 + ... + w_mp_m = 0$
- Determine the angle θ between two vectors $\vec{OP}$ and $\vec{OW}$?
  - We know that
    $$\cos\theta = \frac{\vec{OP}.\vec{OW}}{|\vec{OP}|.|\vec{OW}|}$$
  - Dot product of $\vec{OP}$ and $\vec{OW}$ is $\vec{OP}.\vec{OW} = w_1p_1 + w_2p_2 + ... + w_mp_m$ which is equal to zero. So $\cos\theta = 0$ which implies $\theta = 90°$

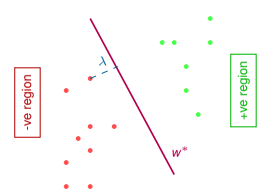
Vector $(w_1, w_2, ..., w_m)$, is normal to hyperplane $w_1x_1 + w_2x_2 + ... + w_mx_m = 0$

Convergence Proof

Satement: If *data is linearly separable*; then the learning rule $w_i = w_i + \alpha y_i x_i$ converges in finitely many steps



Convergence Proof



Update increases w^Tw^* least by $\alpha\lambda$

$$(w + \alpha yx)^Tw^* = w^Tw^* + \alpha yw^{*T}x \geq w^Tw^* + \alpha\lambda$$

$w.w^T$ increases less then α^2

- Let w^* be separating hyperplane
- Make $||w^*|| = 1, \forall_i ||x_i|| \leq 1$
- Let nearest data point be λ distance apart from w^*
- Consider $yw^{*T}x$ it is > 0 for all $x \in D$ with minimum value λ

$$w.w^T \rightsquigarrow (w + \alpha yx).(w + \alpha yx)^T$$

$$= w.w^T + 2\alpha y.w^Tx + \alpha^2y^2x.x^T$$

Update happens only when x is mis classified so $y.w^Tx < 0$. Also $x.x^T \leq 1$ due to scaling and $y^2 = 1$. So, $(w + \alpha yx) \times (w + \alpha yx)^T \leq w.w^T + \alpha^2$

Convergence Proof: After k updates

$$(w + \alpha yx)^Tw^* \geq w^Tw^* + \alpha\lambda \tag{1}$$

$$(w + \alpha yx).(w + \alpha yx)^T \leq w.w^T + \alpha^2 \tag{2}$$

After k updates (when it converges)

$$\begin{aligned} k.\alpha\lambda &\leq w^T.w^* && \text{using eq-1 for } k \text{ number of times} \\ &= ||w^T.w^*|| \leq ||w^T||.||w^*|| && \text{Cauchy Schwarz Inequality} \\ &= ||w|| && \text{as } ||w^*|| = 1 \\ &= \sqrt{W^TW} < \sqrt{k\alpha^2} && \text{using eq-2 for } k \text{ number of times} \\ k.\alpha\lambda &\leq \sqrt{k\alpha^2} && (3) \end{aligned}$$

$\boxed{k \leq 1/\lambda^2}$ Number of update steps is bounded by $1/\lambda^2$

Loss Function

Performance is the closeness of hypothesis with target function

- For example
  - Classification
    $$loss(y, h(x)) = \begin{cases} 1 & \text{if } h(x) \neq y \\ 0 & \text{otherwise} \end{cases}$$
  - Regression
    $$loss(y, h(x)) = \begin{cases} (h(x) - y)^2 & \text{if } h(x) \neq y \\ 0 & \text{otherwise} \end{cases}$$

Issue is that, we can only do Empirical Risk Minimization.

Since only training data is available, on can use this only to be good (minimize risk on empirical data, in hope to reduce generalization risk)

Thank You!

Thank you very much for your attention!

Queries ?