Intro to ML by Tom Mitchelle

ML Murphy

Page No.:
Date: / /

- Classical ML - Statistical
- DL
- Gen AI

Theory
- Mid + End
- Assignments

Lab
- Project
- Presentation

DL Journal for CSE
- ← articles related to opted project
- A/A* main conference — 1 article per member (at least)

Theory (60%.)
- Midsem — 20%.
- Endsem — 30%.
- Assignments - 10%. (Best 2/3)

Lab (40%.)
- Project
  - New/ Unsolved Problem
  - Refractoring the code (already solved but no code)
  - Build new architecture
- 10% SoP
- 5% Presentation
- Phase-I
- Phase-II — 5%. Progress Evaluation
- Phase-III — 20%. Final
  - Project Report (10 pg) (5%.)
  - Presentation (5%.)
  - Code (10%.)

Task T

Performance P

Data/Exp E

⇒ Learner System
- Learner — Takes exp and log → builds model
- Reasoner — works with model
  - Provides sol^n
  - also perf. measure

Creating Learner
- Choose training experience — Features
- Choose target function — Class of functions
- Choose how to repr. target fx^n — Designing of learning algo
- Choose a learning algo to infer target fx^n

chose err fx etc.

Target fx^n → problem type (classification, regression etc.)

↳ $y = f(x)$ $x \rightarrow$ feature, $y \rightarrow$ actual expected o/p

Page No.:
Date: / /

Unsupervised Learning
→ No examples (eg)

Semisupervised Learning
→ few labelled many unlabelled

Reinforcement Learning
→ policy, agent, environment, reward-penalty

Online v Batch
↳ use all to train in one go (offline)
↳ (new+new) + new + ...
↳ incremental

Instance based v Model based (categorization techniques)
↳ measure of similarity ↳ Build a model for prediction

fitness or
Utility metrics — How good
Cost function — How bad

Page No.:
Date: / /

Types of ML Systems:

→ On basis of human supervision
Supervised, Unsupervised, Semi-supervised, Reinforcement
labeled unlabel most unlabelled reward-punishment

→ incremental or learning on fly
Online, Batch
part by part all at once

→ comparing to known data pts or by detecting patterns
instance vs model based
distance model

Supervised Learning
Given: set of input features x_i
target features y_i
set of input-target features
New example

~~Output~~
Prediction
classification - discrete categories
Regression - continuous

Features : attr.
Feature vector : attr vec for numericals
Instance Space X : set of all possible objs that can be defined by features.
Example (x, y)
Concept (c) : subset of X
Target fn : Map each $x \in X$ to $y \in Y$
Training Data : observed egs

Page No.:
Date: / /

Challenges

Insufficient training Data

- overfitting → would perform poorly on new data
- capturing noise rather than features

Non representative Training Data

- inaccurate reflection of distribution and density
- poor generalization

Poor Quality Data

- Data contains errors outliers and missing values
- inaccurate predictions

Irrelevant Features

- including irrelevant features increases complexity and can lead to overfitting
- fits noise rather than meaningful patterns
- more resources required

Overfitting Training Data

- captures noise and random fluctuations too well but does not generalize to new data

Underfitting

- too simplistic to capture the patterns

$E(y - \hat{f})^2 = E[f(x) - \hat{f}(x) + \epsilon]^2 = E[f(x) - \hat{f}(x)]^2 + \text{Var}(\epsilon)$

=

Page No.:
Date: / /

Bias

$y_i - \hat{y}_i$

$\Rightarrow \sum_{i=1}^{N} f(x_i) - \hat{y}_i = \text{EPE}$ (Expected prediction error

True function — prediction function

$= E\left[(f(x_i) - \hat{y}_i)^2 \right]$

$= E[f(x_i)^2] + E[\hat{y}_i^2] - 2E[f(x_i)\hat{y}_i]$

$= f(x_i)^2 + E[\hat{f}(x_i)^2] - 2f(x_i)E[\hat{f}(x_i)]$

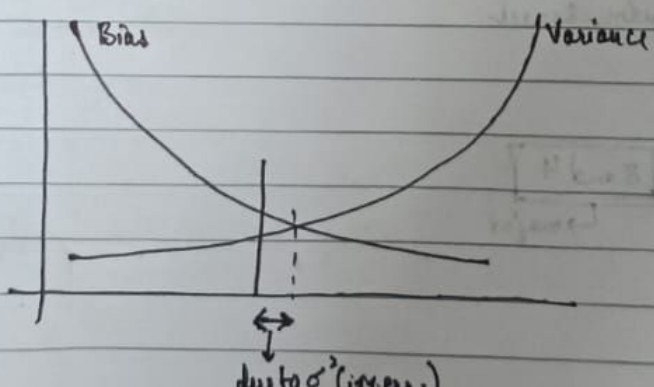
$= E[f(x) - E[\hat{y}(x)]]^2 + E[(\hat{y}(x) - E[\hat{y}(x)])^2] + \sigma^2$

Bias — deviation from avg. prediction (from train) of actual values

Variance — deviation from avg. prediction (from test/val) of pred vals.

Irreducible Error — Error that can not be avoided — due to faults in data collection (eg: faulty/noisy sensors)

Bias Variance Tradeoff



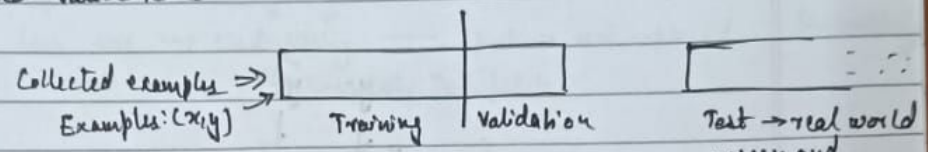
point chosen depends on metric chosen (prec. recall acc.)

same model type (only parameter/hyp diff) → pval
diff model type → acc, prec, recall

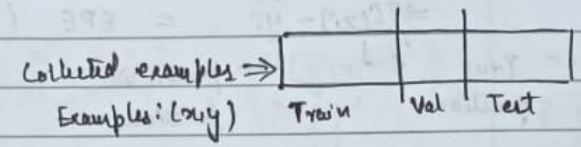
Page No.:
Date: / /

Testing and Validation

Ideal



Experimental



Generalization error = $Error_{test} - Error_{train}$

Gen err less → accurate
more → over/under

Hyper-parameter tuning

Importance of features → parameters
(epochs, batchsize, ...) training → hyperparameters
important for initiating process

→ Manual Search
→ Grid Search
→ Random Search

3 ∉ [1, 2, 3 and 4]
↳ major

No Free Lunch Theorem | Inductive Bias | PAC learning

Page No.:
Date: / /

PAC - Probability Approx. Correct

↑ → Regularization, Feature Selection, Dropouts

low bias high var → Overfitting — noise captured — high complexity
high bias low var → Underfitting — patterns not captured — low complexity

↓ → Complexity, Data points

$$\left(Var = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{M}\sum_{j=1}^{M}\left(y_j^i(x_i) - E[\bar{y}(x_i)]\right)^2\right)$$

multiple instances → multiple model → mean model fixed.

$E[\bar{y}(x_i)] = 5$

$Var = 0.167$

$Bias^2 = 0$

$f(x) = 5$

$x = 2$

$y_1(x) = 4.5$
$y_2(x) = 5$
$y_3(x) = 5.5$

~~can say~~ · can't say over/under without var range
what if var is considered not that high?

Hypothesis Space
→ all possible hypothesis

VC dimensions
→ how effectively we can select hypothesis based on datapts.

generalization bound

~~true err~~

~~train err~~

Inductive bias — Assuming (assumptions) → Effectiveness → PAC Learning
Deduction bias — Inferring (deductions / inferences)

Theoretical framework for evaluation

$P(Error_h \leq \epsilon) \geq 1-\delta$

err threshold 0.5

Err for hypothesis h

δ = confidence

not applying Shattering ← (Vapnik-Chervonenkis) VC ~~learning~~ Theory

If hyp is complex enough but a simpler hyp. has also the same result, pick simpler one.

at · if null is rejected

Tom Mitchelle book

Page No.:
Date: / /

consistent hypothesis - remains same

Eg for find-S

f_1	f_2	f_3		
red	round	small	+	→ add to hypothesis
yellow	long	medium	−	→ check and modify so that - ve are not in hyp
Green	round	medium	+	→ add to hypothesis

Initially

	f_1	f_2	f_3
S_0	$\{\phi_1$	ϕ	$\phi\}$
S_1	{red	round	small}
S_2	: S_1		
S_3	~~{red, green~~	{?, round	?}

⇒ S_3 is c(x)

Version space?

+ve sample → minimal generalization from specific boundary

~~Eg for candidate Elimination~~

-ve sample → minimal specialization from General boundary

Check consistency

Eg

Sky	AirTemp	Humidity	Wind	Water	Forcast	PlaySport
Sunny	Warm	Normal	Strong	Warm	Same	Yes
"	"	High	"	"	"	"
Rainy	Cold	"	"	"	Change	No
Sunny	Warm	"	"	cool	"	Yes

S_0 <ϕ ϕ ϕ ϕ ϕ ϕ> G_0 <? ? ? ? ? ?>

S_1 <Sunny warm normal strong warm same> $G_1 = G_0$

S_2 <Sunny, Warm, ?, strong, Warm, same> $G_2 = G_1$

$S_3 = S_2$ ~~G_3 = <Rainy, ? ? ? ? ?>~~ G_3 = <Sunny

(rule rating v Num rating) and context based rating

Add more emotions-cum

Page No.: 7/22

Date: / Today

Justify role in SoP

$n \rightarrow$ no. of data points.

$VC(H) \rightarrow$ VC dim of hyp space

Generalization Bound

$$Err_{true} \leq Error_{Emp.} + \sqrt{\frac{VC(H)\log(n)}{n}}$$

Err_{true} ↓ test error

$Error_{Emp.}$ ↓ training error

If $VC(H) \rightarrow$ high $\rightarrow E_{true} \rightarrow$ high

If $n \rightarrow$ high $\rightarrow E_{true} \rightarrow$ low

SoP

- PS
- Related Work
- Solution – Flow diagram with phase divisions and work-results div

Concept Learning

→ Binary → true = concept

⟹ Find-S algorithm

⟹ Candidate Elimination Algorithm

Dataset $\mathcal{D} := \{(x_1, y_1), (x_2, y_2), \cdots (x_n, y_n)\}$ $\quad x \in \{1, 1000\}$

$y \in \{True, false\}$

$y_i = c(x_i)$

$\Rightarrow \mathcal{D} = \{(x_1, c(x_1)), (x_2, c(x_2)), \cdots\}$

⇓

need to find $h(x) = c(x) \;\forall\; x \in X$

If all have true for an attr → keep
else not-consistent

$S_3 = S_2$ ⇒ $G_3 = \langle \text{Sunny}, ?, ?, ?, ?, ? \rangle$ (consistent (c)) | $\langle ?, \text{warm}, ?, ?, ?, ? \rangle$ (consistent) | $\langle ?, ?, \text{Normal}, ?, ?, ? \rangle$ (not consistent)

$\langle ?, ?, ?, \text{Cool}, ?, ? \rangle$ (not-c) | $\langle ?, ?, ?, ?, \text{Same} \rangle$ (not-c)

⇒ $S_4 = \langle \text{sunny}, \text{warm}, ?, \text{strong}, ?, ? \rangle$ $G_4 = G_3$

Version space ⟶ space having all possible hypotheses

$\langle \text{Sunny} ? ? \text{strong} ? ? \rangle$
$\langle \text{Sunny warm} ? ? ? ? \rangle$ → $|V| = 3$
$\langle ? \text{ warm } ? \text{ strong } ? ? \rangle$

PS – why are we not taking 3 and ⊥ in each hyp

Drawbacks → not reliable for larger datasets
→ order of hypothesis matters

Decision Tree

- root
- internal node
- leaves

⇒ for supervised learning
⇒ recursively partition the training data into homogeneous regions

for classification

→ less time taking and less resource heavy in test ~~time~~

Page No.:
Date: / /

→ Size and shape
  → internal and leaf node #
  → branching factor
  → depth
  → cross validation to test size & shape

→ At leaf
  → Give constant o/p
  → use KNN
  → use other supervised learning model

⇒ How to split at Internal Nodes
→ split should result in as pure groups as possible
→ for classification, entropy is a measure
  ↳ low entropy → high purity

Techniques to split

→ optimality depends on test / case-by case basis
→ greedy approach is good in some cases

→ information gain → informative rules ?

Leaves should give homogenous / informative o/p

→ highest information gain → root

Page No.:
Date: / /

Entropy and Information Gain

S → set of labelled inputs S from C classes & p_c as fraction of c class inputs

Entropy $H(S) = -\sum_{c \in C} p_c \log_{p_c}$

S → split → S_1, S_2 (disjoint

Reduction in Entropy after split ↑ ↓ Information gain

$$IG = H(S) - \frac{|S_1|}{|S|} H(S_1) - \frac{|S_2|}{|S|} H(S_2)$$

HS → +ve, -ve classes in C

$\Rightarrow H(S) - (+ve\, p_c) \log(+ve\, p_c) - (-ve\, p_c) \log(-ve\, p_c)$

~~outlook~~ ↳

~~0.96~~ 0.94

$S_g =$ S_{sunny} 4_{oc} S_{rain}

3− 2+ 4+ 0− 3+ 2−

$H(S_s) = -\left(\frac{3}{5}\log\frac{3}{5} + \frac{2}{5}\log\frac{2}{5}\right) = \frac{1}{5}\left(3\log 3 + 2\log 2 - 5\log 5\right)$

$= -\left(3\log 3 + 2 - 5\log 5\right)$

$H(S_c) = -(\log 1) = 0$

$H(S_{rain}) = -\left(\frac{3}{5}\log\frac{3}{5} + \frac{2}{5}\log\frac{2}{5}\right) = -(3\log 3 + 2 - 5\log 5)$

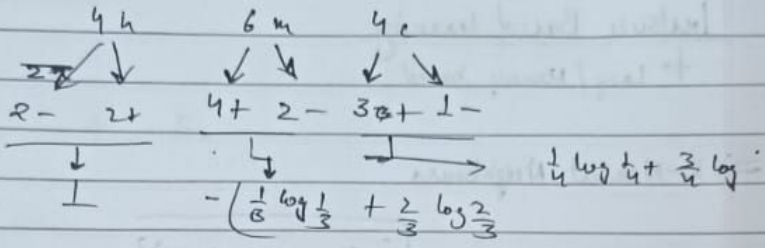
$IG = 0.94 - \frac{5}{14}(\,) - \frac{5}{14} = 0.94 - \frac{5}{14}\left[-(6\log 3 + 2 - 10\log 5)\right]$

patience value

Page No.:
Date: / /

Breiman Leo Friedman, JH, OH Olshen

temperature



When to stop

→ overfit
→ use entropy
→ covered all features

~~decision~~ Avoiding Overfitting in DTs

→ decision stump
→ pre pruning and post pruning → based on purity / entropy
→ feature selection

Notes

→ Gini index instead of IG
→ variance can be used to access purity
→ For real valued features → can use Tests based on thresholding feat. values

---

Decision Tree for Regression

Page No.:
Date: / /

Instance Based Learning

Tom Mitchell

+ Lazy / Memory based

⇒ K-Nearest Neighbours

$$d(x_i, x_j) = \sqrt{\sum_{r=1}^{n} (a_r(x_i) - a_r(x_j))^2} \qquad a_r \to r^{th} \text{ feature}$$

Voronoi Diagram –

⇒ Training Algorithm

For each training algo $\langle x, f(x) \rangle$ add the example to the list of training examples

⇒ Classification Algorithm

$$\hat{f}(x_q) \leftarrow \underset{v \in V}{\operatorname{argmax}} \sum_{i=1}^{k} \delta(v, f(x_i))$$

$$\delta(a,b) = \begin{cases} 1 & \text{if } a=b \\ 0 & \text{o/w} \end{cases}$$

⇒ For regression

$$\hat{f}(x_q) = \frac{\sum_{i=1}^{k} f(x_i)}{k}$$

→ For ~~wrong~~ distance weighted, mult $w_i = \frac{1}{d(x_q, x_i)^2}$ to δ (class) or $f(x_i)$ (regression)

Homo and hetero scedasticity

SVM

Essential Maths for ML
CS.299 Stanford

Page No.:
Date: / /

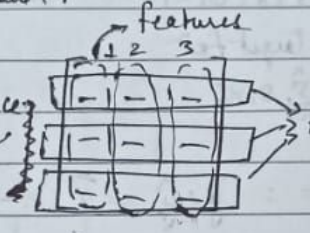
leftnull space ?

Linear Regression

Stat → ISL
Prob → Murphy
Interp → Hands-On

PREREQUISITES

→ Determinant
→ Inversion → Gives opp. vector
→ Vector Space →
→ Linear Transformation
→ Null space (Rank Nullity Thm) (relate with $Ax = 0$) (Gives calc. relation)
→ Metric Space and Normed space
→ Norm $l_1, l_2, \ldots l_\infty$
→ Convex function $\cup$ and Concave $\cap$
→ Multicollinearity
→ Inner product and Inner prod space
↓ dist b/w vectors

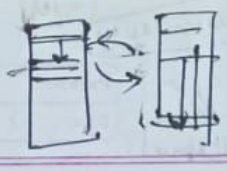


→ Orthogonal → Inner product is zero
→ orthogonal complement → If W is inner product space then $W^\perp$ → if elements of $W^\perp$ are orthogonal to every element of W

$$W^\perp = \{ v \in V \mid \langle w, v \rangle = 0 \ \forall w \in W \}$$

→ Eigenvalue, Eigenvector → search in limited space

→ When to take mean, median, mode
→ IQR, variance range → outlier detection
→ Random Variable
→ distribution
→ pdf, pmf (interpretation of PDF ($p(X=4) = 10.2$))
→ covariance



Page No.:
Date: / /

Linear Regression

Why linear

$y = \theta_1 x_1 + \theta_0 x_0$ ← 1

θ_0 → bias (inductive)

θ_1 → linearity

(for 1 feature)

Hypothesis → $\hat{y} = \theta_0 x_0 + \theta_1 x_1 + \cdots + \theta_n x_n$ (for n-dim / n-featured)

Prediction / finding target fx^n

$\Rightarrow \hat{y} = \sum_{i=0}^{n} \theta_i x_i$

$J(\theta)$ cost fx^n : $y - \hat{y} = \vec{r}$ — (i) (residuals)

$\frac{1}{2} \sum_{i=0}^{n} (\hat{y}_i - y_i)^2$ — (ii) $\equiv r = \underset{\theta}{\operatorname{argmin}} r = \frac{1}{2} \Sigma (\theta_i x_i - y_i)^2$

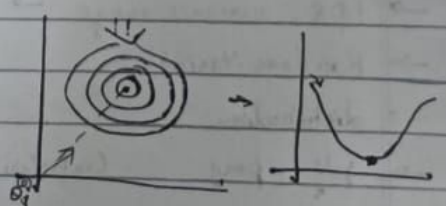
$\frac{1}{n} \Sigma (\hat{y}_i - y_i)^2$ — (iii)

$\frac{1}{n} \Sigma \sqrt{(\hat{y}_i - y_i)^2}$ — (iv)

$h_\theta(x) = \sum_{i=0}^{n} \theta_i x_i$

Goal: $\arg\min_{\theta} J(\theta) \rightarrow \hat{\theta}$ (optimal value of θ for our hypothesis)

If $x \in \mathbb{R}^D$ $\theta \in \mathbb{R}^{d+1}$



How to get the min

↳ Gradient Descent Approach

$\theta^{(0)} =$ init

for all $j \in \{1, \cdots d+1\}$

$\theta_j^{(i+1)} = \theta_j^{(i)} - \alpha \frac{\partial J(\theta)}{\partial \theta_j}$ $\quad i = [0, d]$

Page No.:
Date: / /

$$J(\theta) = \frac{1}{2} \sum (\theta x - y)^T (\theta x - y)$$

Goal $\nabla J(\theta) = 0$

$$\Rightarrow \quad \frac{1}{2} \sum (\theta x - y)^T (\theta x - y) = 0$$

$$\Rightarrow (\theta x - y)^T (\theta x - y) = 0$$

$$\Rightarrow (\theta x)^T \theta x - (\theta x)^T y - y^T (\theta x) + y^T y = 0$$

$$\Rightarrow$$

verify

$$J(\theta) = \frac{1}{2} \sum_{i=1}^{n} (\theta^T x^i - y^i)^2$$

$x = R^{n \times d+1}$ (considering θ_0, x_0)

$\theta = R^{d+1}$

$y = R^n$

$$J(\theta) = \frac{1}{2} \sum_{i=1}^{n} (x\theta - y)^T (x\theta - y)$$

$$= \frac{1}{2} \left[(x\theta)^T x\theta - y^T x\theta - (x\theta)^T y + y^T y \right]$$

$$= \frac{1}{2} \left[\theta^T (x^T x) \theta - 2\theta^T (x^T y) + y^T y \right]$$

$$\Rightarrow \frac{1}{2} \theta^T \left[2 (x^T x) \theta - 2 (x^T y) \right] = 0$$

$$\Rightarrow$$

$$\Rightarrow \theta = \left((x^T x)^{-1} x^T y \right)$$

$(x^T x)\, \theta = (x^T y) \Rightarrow$ Normal Eq^n

Closed form – all given data
GD → real time data.

Page No.:
Date: / /

repeat till convergence
- No change in θ → $\|\theta^{(t)} - \theta^{(t-1)}\|$
- No change in $J(\theta)$ → $\|J_\theta^{(t)} - J_\theta^{(t-1)}\|$
- If $\frac{\partial J(\theta)}{\partial \theta_j}$ is too small

$$\Rightarrow \theta^{t+1} = \theta^t - \alpha \nabla_\theta J(\theta^t)$$
$$= \theta^t - \alpha \nabla_\theta \left[\frac{1}{2}\sum_{i=1}^{n} (h_\theta(x^i) - y^i)^2\right]$$
$$= \theta^t - \alpha \nabla_\theta \left[\frac{1}{2}\sum_{i=1}^{n} (\theta^T x^i - y^i)^2\right]$$
$$\Rightarrow \theta^{t+1} = \theta^t - \alpha \left[\sum_{i=1}^{n} (\theta^T x^i - y^i) x^i\right]$$

θ = Init → BGD

for i in n_samples: (θ = Init) → SGD
  for j in d_features:
    $\theta_j \leftarrow \nabla_{\theta_j} J(\theta_j)$ — SGD
    $acc(\nabla_{\theta_j} J(\theta_j))$ → accumulate BGD
$\theta \in acc(\nabla_{\theta_j} J(\theta_j))$ — BGD

Stochastic v Batch

- BGD converges faster but consumes more mem
- SGD is less resource heavy but converges later
- miniBatch → Best of both worlds → sufficient resources and decent compute time required.

→ ClosedForm Solution

$$J(\theta) = \frac{1}{2}\sum_{i=1}^{n} (\theta^T x^i - y^i)^2$$

$\theta \in \mathbb{R}^{n_d+1}$
$x \in \mathbb{R}^{n_d+1}$
$y \in \mathbb{R}^n$

Page No.:
Date: / /

→ Assumption
⇒ X^TX is invertible ⇒ non-multicollinearity ⇒ linearly independent feature set
⇒ iids

Probability v Likelihood
↳ known value, unknown parameter
↳ known parameter, known distribution

$$y^{(i)} = \theta^T x^{(i)} + \epsilon^{(i)} \qquad \epsilon^{(i)} \sim N(0, \sigma^2)$$
$$i \in iid$$

$$\epsilon^{(i)} = y^{(i)} - \theta^T x^{(i)} \sim N(0, \sigma^2)$$

$$\Rightarrow y^{(i)} \sim N\left(\mu(\theta^T x^{(i)}), \sigma^2\right)$$

$$p(y^i | x^i, \theta) \Rightarrow pdf_N = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(y^{(i)} - \mu(\theta^T x^{(i)})\right)^2}{2\sigma^2}\right)$$

$$L(\theta) \Rightarrow \prod_{i=1}^{n} p(y^i | x^i, \theta)$$

$$= \frac{1}{\sqrt{(2\pi\sigma^2)^n}} \exp\left[\frac{-1}{2\sigma^2} \sum_{i=1}^{n} (y^{(i)} - \theta^T x^{(i)})^2\right]$$

$$\log(L(\theta)) = l(\theta) = \frac{-n}{2} \log 2\pi\sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y^{(i)} - \theta^T x^{(i)})^2$$

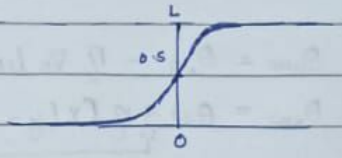
$$l(\theta) = K - \frac{1}{\sigma^2} \underbrace{\sum \frac{1}{2} (y^{(i)} - \theta^T x^{(i)})^2}_{\text{loss function}}$$

$$\Rightarrow \arg\max_\theta l(\theta) = \arg\min_\theta J(\theta)$$

Page No.:
Date: / /

New g choice — logistic / sigmoid

$$g(z) = \frac{1}{1+e^{-z}} \qquad z = \theta^T x$$



$\lim_{z \to -\infty} = 0 \quad \lim_{z \to 0} = 0.5 \quad \lim_{z \to \infty} = 1$

$$\begin{cases} 1 & \text{if } g(z) \geq 0.5 \\ 0 & \text{else} \end{cases}$$

$$h_\theta(x) = g(\theta^T x)$$

$$g(z) = \frac{1}{1+e^{-z}}$$

$$p(y^{(i)}=1 \mid x, \theta) = h_\theta(x)$$
$$p(y^{(i)}=0 \mid x, \theta) = 1 - h_\theta(x)$$

$$g(z)' = g(z)(1-g(z))$$

$$p(y \mid x, \theta) = (h_\theta(x))^{y} \cdot (1 - h_\theta(x))^{(1-y)}$$

$$L(\theta) = \prod_{i=1}^{n} p(y^i \mid x^i, \theta)$$

$$l(\theta) = \sum_{i=1}^{n} y^i \log(h_\theta(x^i)) + (1-y^i) \log(1-h_\theta(x^i))$$

Gradient Descent

$$\nabla_\theta l(\theta) = \sum_{i=1}^{n} \frac{y^i}{h_\theta(x^i)} \cdot \frac{\partial h_\theta(x^i)}{\partial \theta} - \left(\frac{1-y^i}{1-h_\theta(x^i)} \cdot \frac{\partial h_\theta(x^i)}{\partial \theta} \right)$$

For single eg :-

$$= \frac{y^i}{g(\theta^T x)} g'(\theta^T x) x - \frac{(1-y^i)}{1-g(\theta^T x)} g'(\theta^T x) x$$

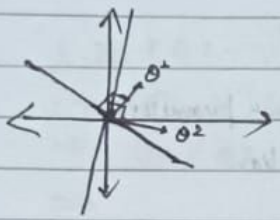
$$\Rightarrow \frac{y}{g(\theta^T x)} g(\theta^T x)(1-g(\theta^T x)) x + \frac{(1-y)(-1)}{1-g(\theta^T x)} g(\theta^T x)(1-g(\theta^T x)) x$$

$$\Rightarrow x\left[y(1-g(\theta^T x)) - (1-y) g(\theta^T x) \right] = x\left[y - y \cdot g(\theta^T x) - g(\theta^T x) + y g(\theta^T x) \right]$$

Page No.:
Date: / /

Perceptron Algorithm

$y^{(i)} = g(\theta^T x^{(i)})$ $\quad y^{(i)} \in \{0, 1\}$



$g \rightarrow$ originally $\begin{cases} 1 & \theta^T x > 0 \\ 0 & o/w \end{cases}$

$g \rightarrow$ better choice $\frac{1}{1 + e^{-\theta^T x}}$ (logistic)

Prereq.

→ Normal Vector

→ Dataset

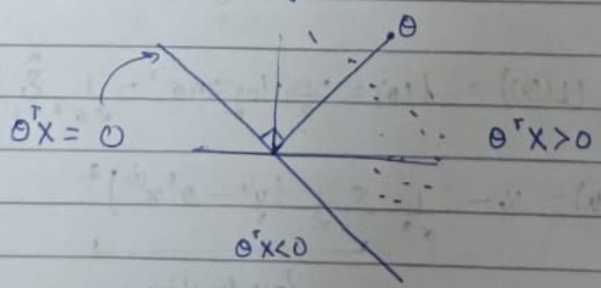
$y = \theta_0 + \theta_1 x$

$Y = \theta^T X$] → Continuous valued / regression

$Y = g(\theta^T X)$

→ g = step function

$\begin{cases} 1 & g() > 0 \\ 0 & g() \leq 0 \end{cases}$



→ step fx^n is very dependent on θ

Higher order norms are not that beneficial

Page No.:
Date: / /

$\Rightarrow \theta_0$ is not regularized

Ridge cost fn

$$J(\theta) = MSE(\theta) + \alpha \cdot \frac{1}{2} \cdot \sum_{i=1}^{n} \theta_i^2$$

Lasso Regression

Least Absolute shrinkage and Selection operator

$$J(\theta) = MSE(\theta) + \alpha \cdot \sum_{i=1}^{n} |\theta_i|$$

Elastic Net

$$J(\theta) = MSE(\theta) + \underbrace{r \cdot \alpha \cdot \sum_{i=1}^{n} |\theta_i|}_{\text{lasso}} + \underbrace{\frac{1-r}{2} \alpha \sum_{i=1}^{n} \theta_i^2}_{\text{ridge}}$$

hyperparams

→ How and why does <which norm makes 0 and lower wt> do what it does (and Elastic net)

→ Create the diagram for Elastic Net, l_1 and l_2

→ Why logistic function ? (for classification) (why not others)
$\frac{1}{1+e^{-x}}$ (Hint: exponential eq^n family)

→ What if No IID in loss functions

→ One SVM question

Newton method → more acc/optimal than GD

Swin-pres
Swin-Q/A
Tue/Fri 11:00

Page No.:
Date: / /

PS, ref article, Flow diag / arch

$$\nabla_\theta = x(y - g(\theta^T x))$$

$$\nabla_\theta = -x(y - g(\theta^T x))$$

$$\Rightarrow \quad \theta_{new} = \theta_{old} - \eta \nabla_\theta l(\theta)$$

$$\Rightarrow \quad \theta_{new} = \theta_{old} + \eta [x(y - g(\theta_{old}^T x))]$$

→ to modify impact on positives

Newton-Raphson approach → Where benefit
→ Where loss

=

Regularizations

→ Good way to reduce anything (reduce the complexity) {

~~dt~~

f_1 dof − 3

⇒ f_1, $f_1 f_1$, $f_1 f_1 f_1$ → 3 features

Lin Reg (f_1, f_1^2, f_1^3)

1 2

12 1^2 2^2 $1^2 2$ 12^2 ||

Ridge Regression

→ reg term $\alpha \cdot \sum_{i=1}^{n} \theta_i^2$ added to cost function

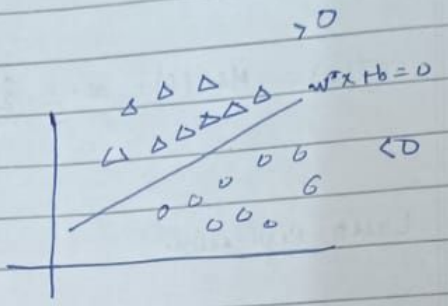
→ forces the learning algo to keep min wts.

primal solution / dual solution | primal solution

Page No.:
Date: / /

SVM

Confidence?



$w^T x + b \geq 1$

$w^T x + b \leq -1$

$w^T x + b = 0$

$y_i (w^T x) \geq 1$

obj fun → $\left[\frac{y_i |w^T x_i + b|}{\|w\|} \right]$

1 - y

y

$\gamma = \min_i \left(\text{Distance (pt}_i, \text{hyperplane)} \right)$

obj ⇒ max γ constraint: ~~optimal~~ $(w^T x + b)$.

⇒ $\underset{w^* b^*}{\text{argmax}}\ \gamma$ constraint: $(w^T x + b)$.

KKT, Lagrangian multiplier ⇒ GD won't work

one idea is ⇒ argmax $\gamma + (w^T x + b)$ → computationally efficient

Final obj → maximize the min dist (pt, hyperplane / dec. surface)

$$\Rightarrow \text{dist}(x_i, w^T x_i + b) = \frac{|w^T x_i + b|}{\|w\|}$$

$$\Rightarrow \text{dist}(x_i^+, w^T x_i^+ + b) = \frac{|w^T x_i^+ + b|}{\|w\|}$$

$$\Rightarrow \text{dist}(x_i^-, w^T x_i^- + b) = \frac{|w^T x_i^- + b|}{\|w\|}$$

Page No.:
Date: / /

⇒ Lagrangian method

$$L(w, b, \alpha) = \underbrace{\frac{1}{2}\|w\|^2}_{\min} + \underbrace{\sum_{i=1}^{n} \alpha_i \left(1 - y_i(w^T x_i + b)\right)}_{\max} \quad \alpha \geq 0$$

→ primal form

~~$\frac{\partial L}{\partial w} = \frac{1}{2} \cdot 2\|w\|_2 + \sum_{i=1}^{n} \alpha_i (1 - y_i($~~

$$\frac{\partial L}{\partial b} = \Sigma \alpha_i y_i = 0$$

$$\frac{\partial L}{\partial w} = \bar{w} - \Sigma \alpha_i y_i x_i = 0$$

max first min later — difficult (?) to compute (why?)

or

max later min first

$$\max_{\alpha \geq 0} \left(\min_{w,b} \left(\frac{1}{2}\|w\|^2 + \sum_{i=1}^{n} \alpha_i \left(1 - y_i(w^T x + b)\right) \right) \right)$$

— verified —

$$\frac{\partial L}{\partial w} = w - \sum_{i=1}^{n} \alpha_i y_i x_i = 0$$

$$\frac{\partial L}{\partial b} = \sum_{i=1}^{n} \alpha_i y_i = 0$$

VERIFY

$\frac{\partial L}{\partial w} = 0$ ~~(at last)~~

$\Rightarrow \frac{\partial L}{\partial w} \Rightarrow$ ~~$\frac{1}{2} \cdot 2\|w\|$~~ $= \sum_{i=1}^{n} \alpha_i y_i x_i = w$

$\frac{\partial L}{\partial b} \Rightarrow \Sigma \alpha_i y_i = 0$

$\frac{\partial L}{\partial b} = 0$ ~~(at last)~~

$$\gamma_{max} = \max_{\alpha \geq 0} \left(\frac{1}{2}\|w\|^2 + \Sigma \alpha_i (1 - y_i(w^T x_i + b)) \right)$$

$$= \max_{\alpha \geq 0} \left(\frac{1}{2}\|w\|^2 + \Sigma \alpha_i - (\Sigma \alpha_i y_i x_i) w^T - (\Sigma \alpha_i y_i) b \right)$$

$$= \quad \rightarrow 0$$

$$L = \Sigma \alpha_i - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j (x_i^T x_j) = w(\alpha) \quad (?)$$

1 2 3 $\frac{1}{1}$ $\frac{1}{2}$ $\frac{1}{3}$

0·1 0·2 $\frac{1}{0.1}$ $\frac{1}{0.2}$

Page No.:
Date: / /

$y_i(w^T x_i + b) \geq 1$ — hard margin (no misclassification)

for +ve $y_i = 1$

$\Rightarrow w^T x_i + b \geq 1$

for −ve $y_i = -1$

$\Rightarrow w^T x_i + b \leq -1$

$\Rightarrow$ dist(xi, hyperplane) $= \frac{y_i |w^T x_i + b|}{\|w\|}$ — verify from here —

$\Rightarrow \min_i \left(\text{dist}(x_i, \text{hyperplane}) \right) = \min_i \hat{\gamma}_i$

$\Rightarrow \gamma = \max(\hat{\gamma}_i)$ $\gamma \rightarrow$ margin

functional margin: $y_i(w^T x_i + b) = \text{dot}(\text{datapt}, \text{hyperplane})$ — sensitive to vals of w and b

geometric margin: $\frac{y_i(w^T x_i + b)}{\|w\|_2}$

$\Downarrow$

1 (? why ?) [verify]

$\Rightarrow \gamma = \max\left(\frac{1}{\|w\|_2}\right)$

$= \min(\|w\|_2)$

$\Rightarrow \left[\gamma = \min\left(\frac{1}{2}\|w\|_2^2\right) ; \quad y_i(w^T x_i + b) \geq 1 \right]$ (for better computation)

Margin-width → controlled by C → determines under/over

Page No.:
Date: / /

$$\mathcal{L} = f(x) + \sum_i \lambda_i g_i(x) + \sum_j \mu_j h_j(x)$$

New obj → $\max_{\alpha} \; w(\alpha)$

$$Q = y_i y_j (x^T x_j)$$

$$\left[\max_{\alpha} \; w(\alpha) = 1^T \alpha - \frac{1}{2} \alpha^T Q \alpha\right] \Rightarrow \text{apply KKT if conditions are True}$$

KKT conditions

- Primal Feasibility $\quad y_i(w^T x_i + b) \geq 1$
- Dual Feasibility $\quad \alpha_i \geq 0$
- Stationarity $\quad w = \sum \alpha_i y_i x_i \;;\; \sum \alpha_i y_i = 0$
- Slackness $\quad \alpha_i \cdot (y_i (w^T x_i + b) - 1) = 0$

→ applying KKT, we get

$$f(x) = \text{sign}\left(\sum_i^n \alpha_i^* y_i (x_i^T x) + b\right)$$

For noisy data,

$$y_i(w^T x_i + b) \geq 1 - \epsilon \qquad \epsilon > 0$$

Soft Margin (may be misclassified)

⇒ Final objective

$$\min_{w,b,\epsilon} \; \frac{1}{2}\|w\|^2 + C\sum_i \epsilon_i \quad \text{s.t.} \quad y_i(w^T x_i + b) \geq 1 - \epsilon_i$$

$$\min_{w,b,\epsilon,\mu,\alpha} \; \frac{1}{2}\|w\|^2 + \mu_j\left(C - \sum_j \epsilon_i\right) + \alpha_i\left(1 - \epsilon_i - y_i(w^T x_i + b)\right)$$

→ verify

$$\frac{\partial \mathcal{L}}{\partial w} \quad \frac{\partial \mathcal{L}}{\partial b} \quad \frac{\partial \mathcal{L}}{\partial \epsilon}$$

Multiclass SVM | One V. | One/KC / all/K

Page No.:
Date: / /

SVM is by default only for binary classification

$$\max_{\alpha} \; w(\alpha) = \sum_i \alpha_i - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j (x_i^T x_j)$$

$$\text{s.t } 0 \le \alpha_i \le C, \quad \sum_i^n \alpha_i y_i = 0$$

if $\alpha_i = 0 \Rightarrow$ data pts are not support vector

if $\alpha_i = C \Rightarrow$ data pts are inside the margin / misclassified

if $0 < \alpha_i < C \Rightarrow$ on the margin

SVM for Non-Linear data

⟹ Kernel trick → project from lower to higher dimension
d H

$H \gg d$

⟹ polynomials in SVM $y_i(w^T x_i + b) \Rightarrow y_i(w^T x + w_i^T z_i + w^T x_i^3 + b) \ge 1$

↳ Very computationally expensive

⇒ Kernel Trick → assuming data is linearly separable in higher dimension

↳ How to project in higher dimension

Kernel functions

↳ Gram Schmidt normalization

↳ Gaussian

↳ Exponential

→ Kernel functions must satisfy M---- condition

(x_i, x_j) → positive semidefinite ~~$A^T A \ge 0$~~

for every datapoint

poly → $K(x_i, x_j) = (x_i^T x_j + c)^d$

rbf / Gaussian / Sigmoid

MSQ

are have done
No place to show

Page No.:
Date: / /

MCQ
- Multiple correct
- fill in the blanks
- numericals → scenario based
- small answers

⇒ Ensemble Learning
- Wisdom of the crowd
  - ↳ aggregation of multiple outputs

Base Classifiers
- Different algorithms
- Different hyper-parameters
- Different representations (Hypothesis)
- Different training set

→ Why → balance bias - variance tradeoff

→ eg → hard voting classifier (eg: majority vote) and soft voting classifier (eg: (aggregate of probabilities)) (hard, soft → voting mech.)

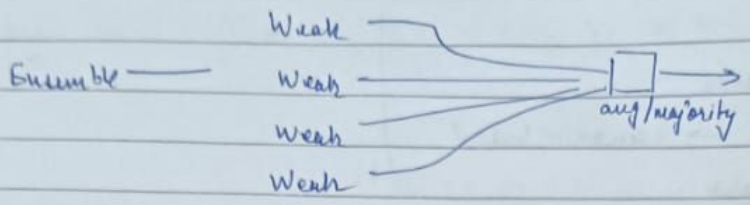
Bagging and Pasting

⇒ Weak Model → Close to 50% probabilities
- → Bootstrapping → Picking random samples from data samples for training (can be in feature level also)
- → Bagging — with replacement
- → Pasting — without replacement

→ how 63? → $n \to \infty$

→ Why not strong instead of weak

Page No.:
Date: / /



Weak ⩾ 50%.

Bag

Dataset → Bagging with replacement (Bootstrapping)

→ Pasting without replacement

36.7% left

OOB

63% ~~picked~~ picked

Boosting → weighted sampling

in dec tree, ~~it~~ it is more prone to overfitting so bagging reduces variance

$$Err = (Bias)^2 + Var + G$$

error prob → $P < 0.5$

if $B \uparrow \Rightarrow P(\text{ensemble voting}) = \sum_{k=\lceil \frac{B}{2} \rceil}^{B} \binom{B}{k} p^k (1-p)^{B-k}$

↑ no. of learners

Var ↓

lowest to a point

Not effective on ~~instance based~~ models like linear regression because var is already low

ISLR book

For skewed datasets → Stratified Bootstrap → ensure correct ratio is picked

Random Forests

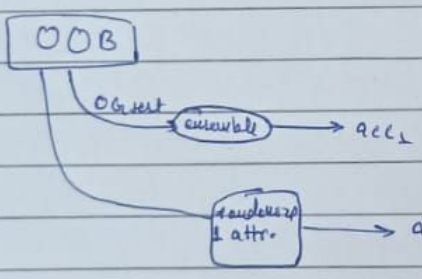
Bagging of datapts and features

Variance Reduction split → such that left subtree var + right subtree var is least

have trees learners trained on a subset of features ($\sqrt{p}$ is a good rule)

⟹ Tree is not optimal here, Ensemble is optimal

Feature importance



if acc_2 is very sharply decreased we can say that feature is important

Can be done using Gini index

→ Slightly less interpretability of RF than DT

→ Hyp → m, B, depth, etc.