(Jan 31 1996)

Exam portion: Moodle - first 4 chapters of textbook.

Exam time 9-11?? Why not 10-12?

$P(x|\omega_1)P(\omega_1) > P(x|\omega_2)P(\omega_2) \Rightarrow \omega_1$.

Here, the decision is made based on apriori probabilities and class conditional probabilities.

6 Feb

* Parametric methods:

- Maximum Likelihood estimate (MLE):

Bayes Theorem: $P(\theta|D) \propto P(D|\theta)P(\theta)$ → joint?
(posterior, likelihood)

$P(X|\theta)$ ← likelihood:

$$l(\theta|X) = P(X|\theta) = \prod_{t=1}^{N} P(x_t|\theta).$$

$X = \{x_t\}_{t=1}^{N}$ ← Draw some (N) data from a known distribution, $P(x|\theta)$.

$x_t \sim P(x|\theta)$. ; $\theta = \begin{bmatrix} \mu \\ \sigma \end{bmatrix}$. $\theta_1 = \begin{bmatrix} \mu_1 \\ \mu_n \end{bmatrix}$; $\theta_2 = \begin{bmatrix} \sigma_1 \\ \sigma_n \end{bmatrix}$.

log likelihood function:

$$L(\theta|X) = \log l(\theta|X)$$
$$= \sum_{t=1}^{N} \log P(x_t|\theta).$$

Bernoulli Distribution: Ber(p). ; Binomial(n, p).

$P(x) = p^x(1-p)^{1-x}$; $x \in \{0, 1\}$.

$$L(p|X) = \log \prod_{t=1}^{N} P(x_t^t)(1-p)^{(1-x_t^t)}.$$

$$= \sum_{t=1}^{N} x_t \log p + (N - \sum_t x_t) \log(1-p).$$

Maximum likelihood Estimate $\hat{p} = \sum_{N} \frac{x_i}{N}$ &

$$\hat{p} = \frac{\sum_i x_i}{N}$$

It is called estimate

* Bias and Variance:

X ~ f_θ Bias of a graph distribution with random

Estimator of θ ← $d = d(X)$ → $\hat{P}(\theta)$

Error = $[d(X) - \theta)^2$

$r(d, \theta)$ = $E[(d(X) - \theta)^2]$

Mean Square Error

Bias of an estimator: $b_\theta(d) = E[d(X)] - \theta$

What does unbiased estimator mean?

For all $b_\theta = 0$, we call it unbiased estimator

→ number of samples

$E[m] = E\left[\frac{\sum_i x_i}{N}\right] = \frac{1}{N} \frac{\sum_i x_i}{N} = \frac{N\mu}{N} = \mu$

↑ concept used similar to

Law of large numbers: When sample size ↑, sample mean ⇝ actual mean

Variance: $Var(m) = Var\left[\frac{\sum_i x_i}{N}\right] = \frac{\sigma^2}{N}$

$Var = E[x^2] - (E[x])^2$

DORA

Mean Square Error:

~~$r(d,\theta) = E[(d(x)-\theta)^2]$~~

$$r(d,\theta) = E[(d(x)-\theta)^2]$$

$$= E[(d - E[d])^2] + (E(d)-\theta)^2$$

Variance + bias.

$$r(d,\theta) = \text{Var}(d) + b_\theta(d)^2$$

Error = Variance + bias^2

13 Feb.

* Dimensionality Reduction:

1. Linear Discriminant Analysis LDA:

Assume we have 2 classes:

Goal: To reduce the dimension while preserving as much as the class discriminatory information as possible.

D dimensional data: $\{x_1, x_2, \ldots, x_n\}$.

$N_1 \rightarrow w_1$ (Class$_1$).

$N_2 \rightarrow w_2$ (Class$_2$).

Get a scalar $y = w^T x$.

Projecting the data points on a line.

Of all possible lines, we select a line that maximizes the separability of scalars, infinite lines are possible, so find the one line among them.

To get a projection vector, define measure of separation

Moon tell me if I could
send up my heart to you(?),
So when I die, which I must do,
could it shine down here with you __/__/__

mean vector: $\mu_i = \frac{1}{N_i} \sum_{x \in w_i} x.$

$$\tilde{\mu}_i = \frac{1}{N_i} \sum_{y \in w_i} y = \frac{1}{N_i} \sum_{x \in w_i} w^T x = w^T \mu_i.$$

Choose the distance b/w projected means as our objective fn.

$$J(w) = |\tilde{\mu}_1 - \tilde{\mu}_2| = w^T(\mu_1 - \mu_2).$$

Variance, too, has to be considered here.

Maximize the difference b/w means normalized by a measure of the within class scatter.

variance ↑

$$\tilde{s}_i^2 = \sum_{y \in w_i} (y - \tilde{\mu}_i)^2.$$

Within class scatter $(\tilde{s}_1^2 + \tilde{s}_2^2)$.
of the projected ~~mean~~ samples.

Fisher linear discriminant is defined as a linear fn $w^T x$ that maximizes the criterion fn:

$$J(w) = \left| \frac{\tilde{\mu}_1 - \tilde{\mu}_2}{\tilde{s}_1^2 - \tilde{s}_2^2} \right|^2.$$

Mean of w_1 & w_2 are close to eachother.

We look for a projection where the examples from the same class projected very close to each other and at the same time, the projected mean are farthest as possible.

Find optimum w^*

$S_W = S_1 + S_2$ Within class scatter.

S_B Between class scatter.

Measure of the scatter in feature space x.

$S_i = \sum_{x \in w_i} (x - \mu_i)(x - \mu_i)^T$

Scatter of projection y — same concept but different space.

$\tilde{S}_i^2 = \sum_{y \in w_i} (y - \tilde{\mu}_i)^2$

$= \sum_{x \in w_i} (w^T x - w^T \mu_i)^2$

$= \sum_{x \in w_i} w^T (x - \mu_i)(x - \mu_i)^T w$

$= w^T S_i w$

$\boxed{\tilde{S}_1^2 + \tilde{S}_2^2 = w^T S_W w}$

Mean: $(\tilde{\mu}_1 - \tilde{\mu}_2)^2 = (w^T \mu_1 - w^T \mu_2)^2$

$= w^T (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T w$

$= w^T S_B w$

$\boxed{J(w) = \frac{w^T S_B w}{w^T S_W w}}$

$\frac{d}{dw}(J(w)) = 0 \Rightarrow \frac{d}{dw}\left[\frac{w^T S_B w}{w^T S_W w}\right] = 0$

$\Rightarrow [w^T S_W w] \frac{d}{dw}[w^T S_B w] - [w^T S_B w] \frac{d}{dw}[w^T S_W w] = 0$

$\Rightarrow [w^T S_W w] 2 S_B w - [w^T S_B w] 2 S_W w = 0$

Scikit learn.

Dividing by $[w^T S_W w]$, we get:

$$\frac{[w^T S_W w]}{[w^T S_W w]} S_B w - \frac{[w^T S_B w]}{[w^T S_W w]} S_W w = 0.$$

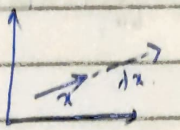
$$S_B w - \frac{[w^T S_B w]}{[w^T S_W w]} S_W w = 0.$$

$$S_B w - \lambda S_W w = 0.$$

$$S_W^{-1} S_B w - \lambda w = 0.$$

$$\rightarrow \boxed{S_W^{-1} S_B w = \lambda w.}$$

Eigen value problem:
$Ax = \lambda x$.
Operator A acts on the vector x to scale x along λ.
No change in direction.



Optimized w : w^*.

$$w^* = \text{argmax} \left[\frac{w^T S_B w}{w^T S_W w} \right] = S_W^{-1} (\mu_1 - \mu_2).$$

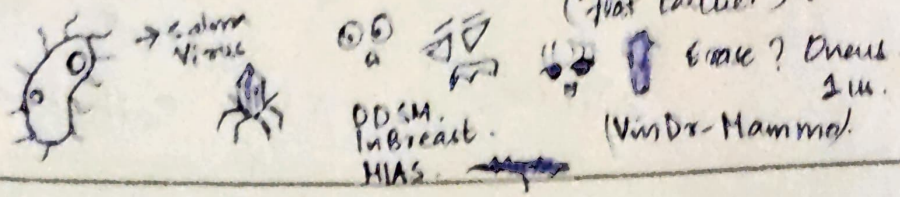
25 Feb.

* Principle Component Analysis:

Linear combination of features that has max. variance.

Scaling data - normalization
Mean $\rightarrow 0$.

Manifold ?? - Generalization of Euclidean space.
- Non linear space.
- Eg: Swiss roll?



5 __/__/__

27 Feb

* Non Linear Dimensionality Reduction by:
- Locally Linear Embedding (LLE)

Making sure the relationship (distance) among the data points are preserved as far as possible.

LLE is a mapping which preserves the neighbourhood.

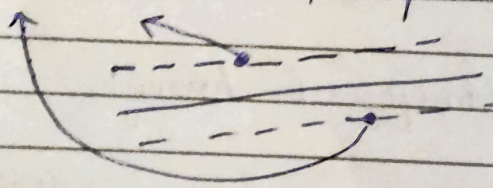
Reconstruction: $\phi(Y) = \sum_i | \vec{Y}_i - \sum_j w_{ij} \vec{Y}_j |^2$ (?).

Note: In higher dimensions, the notion of distance / similarity is different.

4 March

* SVM - Support Vector Machine:
- Supervised.
- Mainly for classification, can also be used for regression.
- For n dimensional data, (n-1) dimensional hyperplane.

Support vectors: vectors / datapoints on the margin.



Hard margin: Margin defined clearly.

Soft Margin Classifier:
Instead of hard margins, we allow misclassification.

Tutor

__/__/__

- Non-linear:
  - for n dimensions, might need to map to (n+1) dimensions to get a linearly separable space.

Kernel tricks:

- RBF:
  $K(x, x') = e^{-\gamma \|x - x'\|^2}$.
  - can handle infinite dimensions.
  - γ - gamma - decides how big/small the margin is.

Optimization? finding the optimal hyperplane.

$z = \phi(x)$

where $z_j = \phi_j(x)$, $j = 1, \dots, k$.

discriminant: $g(z) = w^T z$.

$$g(x) = w^T \phi(x)$$
$$= \sum_{j=1}^{k} w_j \phi_j(x).$$

Replace the inner product of basis functions, $\phi(x^t)^T \phi(x^s)$, by kernel function $k(x^t, x^s)$

$$g(x) = w^T \phi(x) = \sum_t \alpha^t r^t \phi(x^t)^T \phi(x) = \sum_t \alpha^t r^t k(x^t, x).$$

* ~~Exploration & Exploitation:~~

Regression methods ⇒ for regression not a tree

6 March.

* Exploration & Exploitation:
Exploring everything and exploiting the optimal one.

* CART: Classification & Regression Trees:

	CGPA	Interactive	Knowledge	Softskills	Job Offer
1.	≥9	Yes	V good	Good	Yes
2.	≥8	No	Good	Moderate	Yes
3.	≥9	No	Avg	Poor	No
4.	<8	No	Avg.	Good	No
5.	≥8	Yes	Good	Moderate	Yes
6.	≥9	Yes	Good	Moderate	Yes
7.	<8	Yes	Good	Poor	No
8.	≥9	No	VGood	Good	Yes
9.	≥8	Yes	Good	Good	Yes
10.	≥8	Yes.	Avg.	Good	Yes

$$\text{Gini Index}(T) = 1 - \sum_{i=1}^{n} P_i^2$$

where n is the number of labels (Yes, No, ⇒ 2).

$$\text{Gini Index}(T) = 1 - \left[\left(\frac{7}{10}\right)^2 + \left(\frac{3}{10}\right)^2\right] = 1 - \frac{49+9}{100}$$

$$= 1 - \frac{58}{100} = \frac{100-58}{100} = \frac{42}{100} = 0.42.$$

0.42 for whole dataset.
Gini Index for each attribute now:

		Job Yes	Job No.
Gini Index (CGPA) ⇒	≥9	3	1
	≥8	1	0
	<8	0	2

Subsets: $\{\{\geq 9, \geq 8\}, \{< 8\}\}$ A.

$\{\{\geq 8, < 8\}, \{\geq 9\}\}$ B.

$\{\{\geq 9, < 8\}, \{\geq 8\}\}$ C.

Gini Index (T, CGPA wrt $\{\geq 9, \geq 8\}$)

$= 1 - \left[\left(\frac{7}{8}\right)^2 + \left(\frac{1}{8}\right)^2\right] = \frac{64 - 50}{64} = \frac{14}{64} = 0.21875$

Gini Index (T, CGPA $\in \{< 8\}$) $= 1 - \left[\left(\frac{0}{2}\right)^2 + \left(\frac{2}{2}\right)^2\right] = 0$

Combining:

A. Gini Index (T, CGPA $\in \{\{\geq 9, \geq 8\}, \{< 8\}\}$)

$= \frac{|S_1|}{|T|} GI(S_1) + \frac{|S_2|}{|T|} GI(S_2)$

$= \frac{8}{10} \cdot 0.21875 + \frac{2}{10} \cdot 0 = 0.175$

Gini Index (T, CGPA $\in \{\geq 8, < 8\}$) $= 1 - \left[\left(\frac{4}{6}\right)^2 + \left(\frac{2}{6}\right)^2\right] = 1 - \frac{20}{36} = 0.44$

Gini Index (T, CGPA $\in \{\geq 9\}$) $= 1 - \left[\left(\frac{3}{4}\right)^2 + \left(\frac{1}{4}\right)^2\right] = 1 - \frac{10}{16} = 0.375$

B. Gini Index (T, CGPA $\in \{\{\geq 8, < 8\}, \{\geq 9\}\}$)

$= \frac{|S_1|}{|T|} GI(S_1) + \frac{|S_2|}{|T|} GI(S_2)$

$= \frac{6}{10} \cdot 0.44 + \frac{4}{10} \cdot 0.375 = 0.414$

Gini Index (T, CGPA $\in \{\geq 9, \geq 8\}$) $= 1 - \left[\left(\frac{3}{6}\right)^2 + \left(\frac{3}{6}\right)^2\right] = 1 - \frac{18}{36} = 0.5$

Gini Index (T, CGPA $\in \{\geq 8\}$) $= 1 - \left[\left(\frac{4}{4}\right)^2\right] = 0$

C. Gini Index (T, CGPA $\in \{\{\geq 9, \geq 8\}, \{\geq 8\}\}$

$= \frac{|S_1|}{|T|} GI(S_1) + \frac{|S_2|}{|T|} GI(S_2)$

$= \frac{6}{10} 0.5 + \frac{4}{10} \cdot 0 = 0.3$

Gini (CGPA) ≥ 0.175 ✓
≥ 0.24
< 0.3

Compute Δ Gini:
Δ Gini (CGPA) = Gini (T) − Gini (T, CGPA)
= 0.42 − 0.175
= 0.245

Interactive:

	Job Yes	Job No.
Yes	5	1
No	2	2

Gini (T, Interactive ∈ {Yes}) = $1 - \left[\left(\frac{5}{6}\right)^2 + \left(\frac{1}{6}\right)^2\right] = 1 - \frac{26}{36} = 0.28$

Gini (T, Interactive ∈ {No}) = $1 - \left[\left(\frac{2}{4}\right)^2 + \left(\frac{2}{4}\right)^2\right] = 1 - \frac{8}{16} = 0.5$

Gini combined (T, Interactive ∈ {{Yes}, {No}}) = $\frac{|S_1|}{|T|}$ Gini(S_1) + $\frac{|S_2|}{|T|}$ Gini(S_2)

= $\frac{6}{10} \cdot 0.28 + \frac{4}{10} \cdot 0.5$

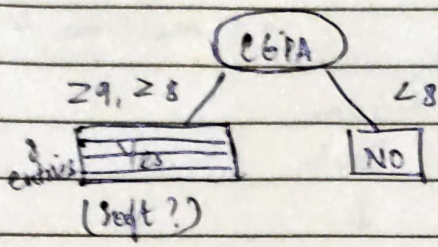
= 0.168 + 0.2 = 0.368.

Δ Gini (Interactive) = Gini (T) − Gini (T, Interactive)
= 0.42 − 0.368
= 0.052.

Using Attribute:
→ CGPA — Gini Index 0.175 — Δ Gini 0.245
Interactive — 0.368 — 0.052
Knowledge — 0.3054 — 0.1146 (?)
→ Soft Skills — 0.175 — 0.245 (?)

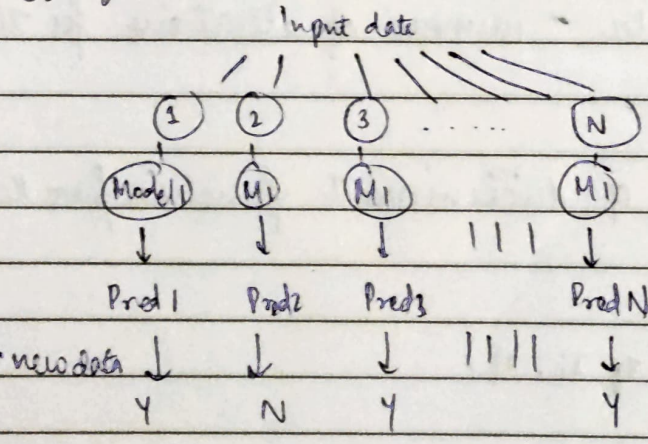
__/__/__

Eg: CGPA vs Soft Skills:



* Ensemble Learning:
- Supervised Machine Learning.
- Combines predictions from multiple models.
- Voting (for classification) / Averaging (for regression).
- Bagging (Bootstrap AGGregation).
- Random forests.
- Boosting.
- Stacked Generalization (Blending).

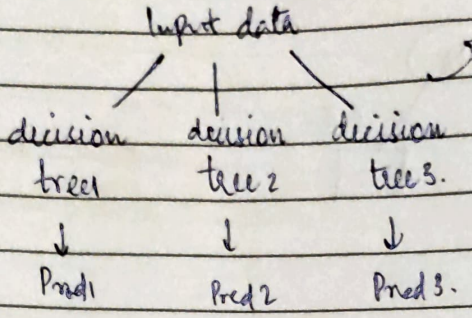
@ Bagging:



Voting: #Y, #N ⇒ Majority wins → classifies.

Thanksgiving!!!
My mook! too cold...! __/__/__
It's dying!! 😭

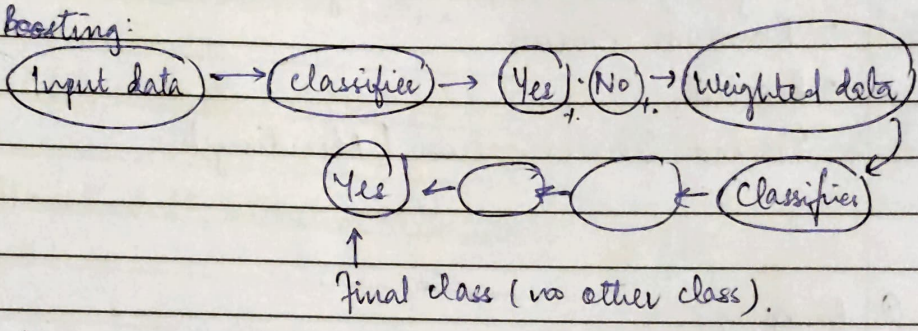
② Random Forests:



Voting / Averaging based on classification / regression.

↑ random forest classifier for final prediction.

③ Boosting:



Hyperparameter - number of iterations for this.

* Derivation of discriminant function from Kernel fn:

$K(x,y)$

$k\{i,j\} = k(x_i, x_j)$

$w = \arg\min \sum_{\{j=1\}}$

[From pdf]. α_i → Lagrange multipliers.

__/__/____

$f(x) = \sum_{i=1}^{n} \alpha_i y_i k(x_i, x) + b$.

wobei b .. bias-term → hyperparameter.