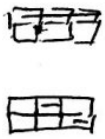
Data Mining 1/24 Lecture 2

Data type: Categorical
Continuous.

Data Processing ⇒ Data Mining ⇒ Postprocessing

OSEMIN: Obtain → Scrub → Explore → Model → Interpret
SQL — Clean → Data Quality — find max/min, mean, distribution — machine learning

Handle missing values:
- List wise deletion
- Pair wise deletion
- Imputation



Quality:
- Missing values.
- Errors / inconsistent values.
- Duplicate values
- Noise / Outliers
  - Noise → typos
  - Outliers → legitimate values but not in range

Exploratory Data Analysis (EDA)

Univariate Analysis : Categorical

Bivariate Analysis : Continuous Vs Continuous.

Scanned with CamScanner

Data Mining 1/29 | Tuesday Lecture 3 (1)

Feature Engineering

columns

Raw Data → Data Cleaning & Exploration → Feature Engineering → Model Building

Pre-processing

- Feature Transformation such as (take log)
- Feature Creation
- Feature Scaling (should be performed inside the cross-validation loop)
  - * Normalization : Scale all data to [0-1]
    $$z = \frac{x - \min(x)}{\max(x) - \min(x)}$$
  - * standardization : Rescale to have a mean of 0 and standard deviation of 1
    $$z = \frac{x_i - \mu}{\sigma}$$

Noise / Outliers

leave them as-is
Deletion
Trans :
Binning
Imputation
Separation : seperately process

Binning / Discretization : transform a cont. attribution into a categorical / nominal attribute (or to a binary, binerization)
eg. k-means, Equal frequency.

Sampling : Reduce the dataset size

Aggregation : data by date => aggregate to by week/month

Dimensionality : The number of attributes in data set
We can think of each row as a d-dimensional point
(eg. show the 4th dimension by color)

Scanned with CamScanner

Curse of Dimensionality

Dimensionality reduction is the process of reducing the number of variables under consideration. There are 2 methods:

- feature selection — by selecting a subset of attributes
- feature extraction

Domain knowledge
Missing values ratio
Low variance filter
High correlation filter
Feature creation

Feature Selection Methods:

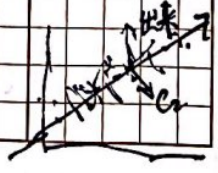
- Filter Methods: Score each feature based on statistical correlation with target value
- Wrapper Methods: Train a model using a subset of features. Based on the results, decide to add or remove features from our subset.
- Embedded Methods: Feature selection occurs as part of the data mining algorithm.

Feature Extraction / Dimensionality Reduction:

Use techniques from Linear algebra to project the data from a high-dimensional space to a low-d space in such away that maximum variance of data is captured.

(PCA) Principle Component Analysis
(SVD) Sigular Value Decomposition
(LDA) Linear Discriminant Analysis.

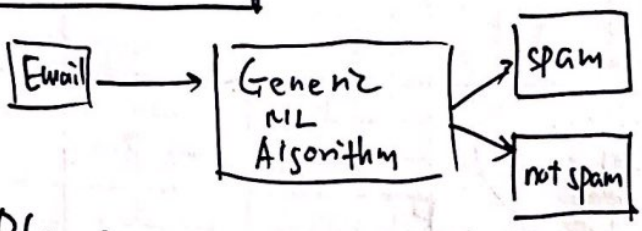
PCA: identifies the best hyperplane to project the data onto



Scanned with CamScanner

Data Mining 4/2 Tue Lecture 3 (2)

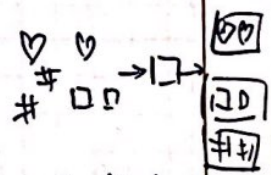
Machine Learning



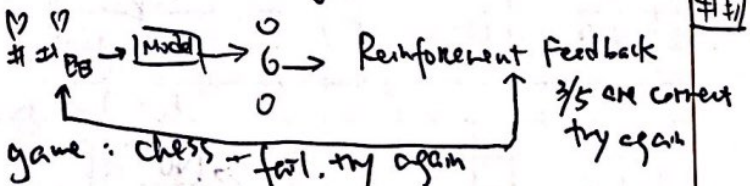
ML Ref AI : ML is a subset of AI

Supervised learning: Labelled training data

Unsupervised learning: unlabelled trang data



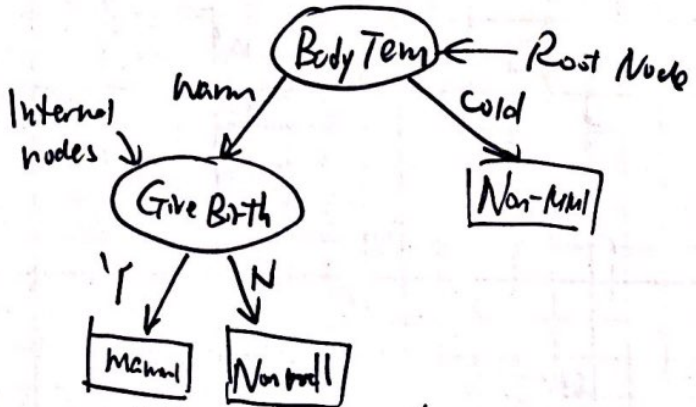
Reinforcement learning:



Classification (Supervised learning)

Assigning.

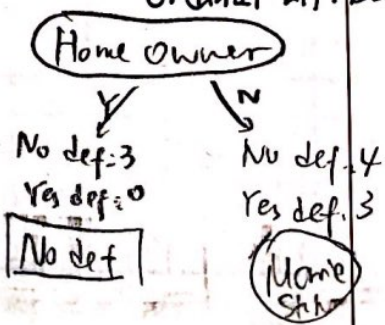
Decision Trees



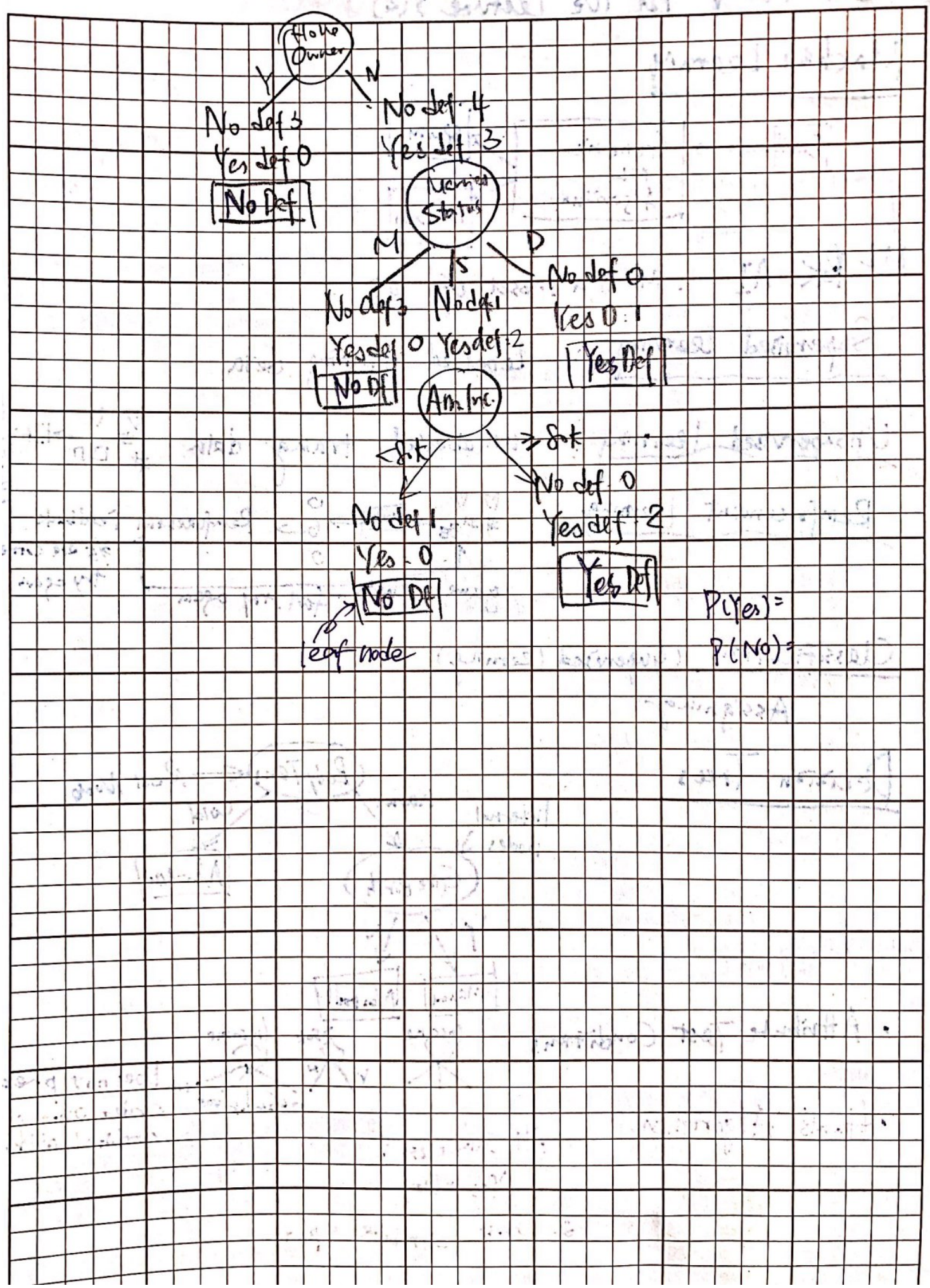
- Attribute Test Conditions

- Hunt's Algorithm

1. All Same class?
   Yes → leaf
2. Choose attribute split



Scanned with CamScanner



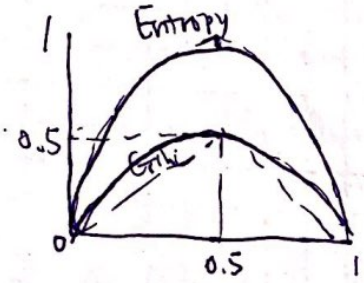
Scanned with CamScanner

Greedy Algorithm : which node is the best?

Selecting the best split.

Example impurity measures:

- Entropy : $E(S) = \sum_{i=1}^{c} -p_i \log_2 p_i$
- Gini : $Gini = 1 - \sum_{i=1}^{c} (p_i)^2$



Comparison among the impurity measures for binary classification

GAIN parent node → child node the GAIN↑, the better split.

$$GAIN_{split} = Impurity(p) - \left(\sum_{i=1}^{n} \frac{n_i}{n} Impurity(i) \right)$$

example impurity of parent: $Gini_{(parent)} = 1 - (\frac{3}{10})^2 - (\frac{7}{10})^2$ (Yes, No) [loan/class example/Default]

$Impurity = 0.42$

★ if split homeowner

Y / N

Def No: 3, Def Yes: 0 — $Gini = 1 - (\frac{0}{3})^2 - (\frac{3}{3})^2 = 0$

Def No: 4, Def Yes: 3 — $Gini = 1 - (\frac{3}{7})^2 - (\frac{4}{7})^2 = 0.489$

$\sum Gini(split) = (\frac{3}{10}) \cdot 0 + (\frac{7}{10}) \cdot 0.489 = 0.342$ (Yes, No)

$GAIN = 0.42 - 0.342 = 0.078$

★ if split on Marital status

S / M / D

Def No: 2, Def Yes: 2 — $Gini = 0.5$

Def No: 4, Def Yes: 0 — $GINI = 0$

Def No: 1, Def Yes: 1 — $GINI = 0.5$

$\sum Impurity \frac{n_i}{n} = (\frac{4}{10}) 0.5 + (\frac{4}{10}) \cdot 0 + (\frac{2}{10}) 0.5 = 0.3$

$GAIN = 0.42 - 0.3 = 0.12$

Scanned with CamScanner

Split on Annual Income ≤ 55K.

0	10
Def N: 0	Def N: 7
Def Y: 0	Def Y: 3
Gini = 0	GINI = $1-(\frac{3}{10})^2-(\frac{7}{10})^2$

Gini: $(\frac{0}{10}) \cdot 0 + (\frac{10}{10}) \cdot 0.42$

$\sum_N$

GAIN = $1-(\frac{3}{10})^2-(\frac{7}{10})^2 - \boxed{\sum} = 0$.

on Annual Income ≤ 65K

1/10 Y	9/10 N
N: 1	N: 6
Y: 0	Y: 3
Gini = 0	GINI = 0.445

$\sum$ Impurity $\frac{n_i}{n}$ = $(\frac{1}{10}) \cdot 0 + (\frac{9}{10}) \cdot 0.445 = 0.4$

GAIN = 0.42 − 0.4 = 0.02

Example: → to choose the least Σ Gini. and split

	Y	3	0
	N	3	4

0.343

① MS
② HO
③ AIncome 90K

Choose the highest GAIN

→ Continuous Variables.

GAIN Ratio

$$\text{GainRATIO} = \frac{\text{GAIN}_{\text{split}}}{\text{SplitINFO}}$$

$$\text{SplitINFO} = -\sum_{i=1}^{k} \frac{n_i}{n} \log_2 \frac{n_i}{n}$$

Scanned with CamScanner

Data Mining 1/31 Thursday Lecture 4(2)

GAIN Ratio continue

example $SI_{(\text{marital status})} = -\frac{4}{10}\log_2(\frac{4}{10}) - \frac{4}{10}\log_2\frac{4}{10} - (\frac{2}{10})\log_2(\frac{2}{10})$

$= 1.52$

$\Rightarrow$ GAINRATIO $= \frac{0.12}{1.52} = 0.079$ for Marital Status Attributes.

$SI_{(\text{Home owner})} = -\frac{3}{10}\log_2\frac{3}{10} - \frac{7}{10}\log_2\frac{7}{10} = 2.88$

$\Rightarrow$ GAINRATIO $= \frac{0.078}{2.88} = 0.089$ for Attribute Home Owner

TIP: If all the split in the same data type (binary), choose GAIN otherwise, choose GainRATIO

Decision Boundaries: use axis-parrel hyperplane to split data space into purer partitions.

Data Fragmentation

Scanned with CamScanner

Data Mining 2/5 Tuesday Lecture 5 | 17

Decision Tree Algorithm

Regression Trees

Practice Problem

$Gini(parent\ Impurity) = 1-(\frac{4}{7})^2-(\frac{3}{7})^2 = 1-\frac{16}{49}-\frac{9}{49} = 0.49$

① if split by Pass all Assignments

Y 4 / N 3

Pass Y : 2 | Pass Y : 2
Pass N : 2 | Pass N : 1

$GINI = 1-(\frac{2}{4})^2-(\frac{2}{4})^2 = 1-\frac{1}{4}-\frac{1}{4} = \frac{1}{2}$

$GINI = 1-(\frac{2}{3})^2-(\frac{1}{3})^2 = 1-\frac{4}{9}-\frac{1}{9} = \frac{4}{9}$

$\Sigma_{GINI(split)} = \frac{4}{7}\cdot\frac{1}{2}+\frac{3}{7}\cdot\frac{4}{9} = 0.476$

$GAIN = 0.49-0.476 = 0.014$

$GAINRATIO = \frac{GAIN}{SI} = \frac{0.014}{-\frac{4}{7}\log_2\frac{4}{7}-\frac{3}{7}\log_2\frac{3}{7}} = 0.019$

2.0	2.5	2.1	3.2	3.3	3.5	3.9
No	No	Yes	Yes	Yes	Yes	No

② if split by GPA ≤ 3.1

Y 3 / N 4

Pass Y : 1 | Pass Y : 3
Pass No : 2 | Pass N : 1

$GINI = 1-(\frac{1}{3})^2-(\frac{2}{3})^2 = 1-\frac{1}{9}-\frac{4}{9} = \frac{4}{9}$

$GINI = 1-(\frac{3}{4})^2-(\frac{1}{4})^2 = 1-\frac{9}{16}-\frac{1}{16} = 1-\frac{10}{16} = \frac{6}{16} = \frac{3}{8}$

$\Sigma_{GINI(split)} = (\frac{3}{7})\cdot\frac{4}{9}+(\frac{4}{7})\cdot\frac{3}{8} = 0.405$

$GAIN = 0.49-0.405 = 0.085$

GPA ≤ 3

Y 2 / N 5

Pass Y : 0 | Pass Y : 4
Pass N : 2 | Pass : 1

$GINI = 1-(\frac{0}{2})^2-(\frac{2}{2})^2 = 0$

$GINI = 1-(\frac{4}{5})^2-(\frac{1}{5})^2 = 1-\frac{16}{25}-\frac{1}{25} = \frac{8}{25}$

$\Sigma_{GINI(split)} = (\frac{2}{7})\cdot 0+\frac{5}{7}\cdot\frac{8}{25} = 0.229$

$GAIN = 0.49-0.229 = 0.261$

$GR = \frac{0.26}{-\frac{2}{7}\log_2\frac{2}{7}-\frac{5}{7}\log_2\frac{5}{7}} = 0.302$

Scanned with CamScanner

(3) If Split On Language

Python 3	C++ 2	Java 2
Pass Y: 1	Pass Y: 2	Pass Y: 1
Pass N: 2	Pass N: 0	Pass N: 1

$GINI = 1-(\frac{1}{3})^2-(\frac{2}{3})^2 = 1-\frac{1}{9}-\frac{4}{9} = \frac{4}{9}$

$GINI = 1-(\frac{2}{2})^2-(\frac{0}{2})^2 = 0$

$GINI = 1-(\frac{1}{2})^2-(\frac{1}{2})^2 = 1-\frac{1}{4}-\frac{1}{4} = \frac{1}{2}$

$\sum GINI(split) = (\frac{3}{7})\cdot\frac{4}{9} + (\frac{2}{7})\cdot 0 + (\frac{2}{7})\cdot\frac{1}{2} = 0.3333$

$GAIN = 0.489 - 0.333 = 0.157$

$GAINRATIO = \frac{GAIN}{SI(Lan)} = \frac{0.157}{-\frac{3}{7}\log_2\frac{3}{7} - \frac{2}{7}\log_2\frac{2}{7} - \frac{2}{7}\log_2\frac{2}{7}}$

$= \frac{0.157}{1.557}$

$= 0.1$

Compare the 3 ways of GAINRATIO

choose ~~Language~~ GPA to split on first.

Scanned with CamScanner

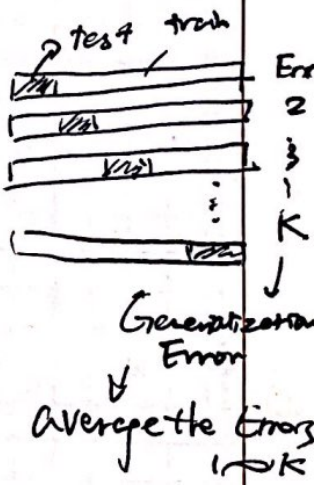
Data Mining 2/5 Tue Lecture 5 (2)

Cross-Validation & Overfitting

Types of Error
- Traing Error in Training Phase
  the percent of misclassification errors on the traing set.
- Testing Error / Generalization Error
  ——————— on the test data.

Partitioning Data to Predict Error

K-Fold Cross-Validation : split data into K folds



Scanned with CamScanner

Data Mining 1/3 Thursday Lecture 6

Overfitting

An over-fitted model: too specific to the training data and will not generalize well to new data.

Pre-pruning: stop grow the tree before it is fully grown.

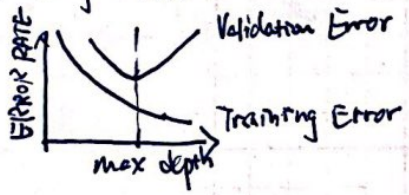
eg. χ^2 (Chi-square) pruning example. $\chi^2 = \sum_{\substack{\text{all classes } i \\ \text{children } j}} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}$

null: you split on random data. correctly.

Post-pruning Trim the fully grown tree from the bottom up.

- Reduced Error Pruning (REP)

Model Selection
=
hyperparameter tuning.



Re ~~select~~ ~~cross validation~~ select ~~test~~ data/val.

Scanned with CamScanner

Data Mining 2/12 Tuesday Lecture 7

Nearest Neighbor Classifiers

Eager Learner

Lazy Learner

K-nearest Neighbors : use euclidian distance, $d = \sqrt{\sum (p_i - q_i)^2}$

Scaling

for nominal attribution: distance = 0 if they same
distance = 1 if different

Voting Schemes

weight ~~vectors~~ factor : $w = \frac{1}{d^2}$

Decision Boundaries

k = 1 , overfitting.

k = 20

⋮ underfitting

⇓

choose the right k : using cross-validation

Instance Reduction Algorithms

In Class Activity -

$d_{12} = \sqrt{1 + (7-7)^2 + (7-4)^2} = \sqrt{1+9} = \sqrt{10}$

$d_{13} = \sqrt{0 + (7-3)^2 + (7-4)^2} = \sqrt{16+9+1} = \sqrt{26} = 5$

$d_{14} = \sqrt{1 + (7-1)^2 + (7-4)^2} = \sqrt{1+36+9+1} = \sqrt{47} = \sqrt{46}$

$d_{23} = \sqrt{1 + 4^2 + 0^2} = \sqrt{1+16} = \sqrt{17}$

$d_{24} = \sqrt{1 + 6^2 + 0^2} = \sqrt{37}$ $\sqrt{37}$

$d_{34} = \sqrt{1 + 2^2 + 0 + 0} = \sqrt{5}$

X X

Scanned with CamScanner

Data Mining 2/14 Thursday Lecture 8

Bayesian Classifiers : Naive Bayes : Bayes Theorem and a "naive" assumption (independent)

$P(A, B) = P(A) * P(B|A)$ dependent

$P(A, B) = P(A) \times P(B)$ Independent.

Bayes Theorem $P(B|A) = \frac{P(A|B) \cdot P(B)}{P(A)}$ ← class probability, conditional

Classification $P(C|X) = \frac{P(X|C)\,P(C)}{P(X)}$ ← Prior

$P(C|X)$ ← posterior

Example:

$P(\text{not def} \mid \{9.5000, \text{own}\}) \propto P(T=10|\text{not}) \times P(S=1000|\text{not}) \times P(\text{own}|\text{not}) \times P(\text{NOT DEF})$

$\propto \frac{4}{6} \times \frac{3}{6} \times \frac{4}{6} \times \frac{6}{10}$

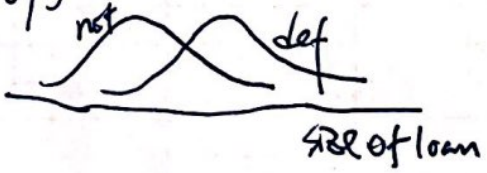
$= 0.1334$

$P(\text{DEF} \mid \{9.5000, \text{own}\}) \propto P(T=10|\text{DEF}) \times P(S=10000|\text{DEF}) \times P(\text{own}|\text{DEF}) \times P(\text{DEF})$

$= \frac{3}{4} \times \frac{4}{4} \times \frac{1}{4} \times \frac{4}{10}$

$= 0.075$

Continuous Attributes: draw the pdf



Laplace smoothing if: Zero out an attribute value we've never seen

Add 1 to the numerator and denominator of the class-conditional prob.

$P(\text{Green}|\text{Apple}) = \frac{1}{4}$ (本来是 $\frac{0}{3}$)

$P(\text{Green} \mid \text{Banana}) = \frac{0}{1}$ 本来 $\xrightarrow{\text{Laplace smoothing}} \frac{1}{2}$ to avoid zeroing the attribute out.

Scanned with CamScanner

Decision Boundaries / Probability Distributions.

Characteristic: Naive Bayes Attributes Independent.

Text Classification

Bag of words model: D_1: Each state has its own laws.

D_2: Every country has its own culture

Class	each has				
D_1	0	1	1	0	0
D_2	-	0	1	-	-

Example. $P(C|X) = \frac{P(X|C) \times P(C)}{P(X)}$

$$P(\text{Spam} | R, H) = P(R|\text{SPAM}) \times P(H|\text{SPAM}) \times P(\text{SPAM})$$

$$= \frac{400}{500} \times \frac{10}{500} \times \frac{500}{1500}$$

$$= 0.0053$$

$$P(\text{Ham} | R, H) = P(R|\text{Ham}) \times P(H|\text{HAM}) \times P(\text{HAM})$$

$$= \frac{25}{1000} \times \frac{50}{1000} \times \frac{1000}{1500}$$

$$= 0.00083$$

Scanned with CamScanner

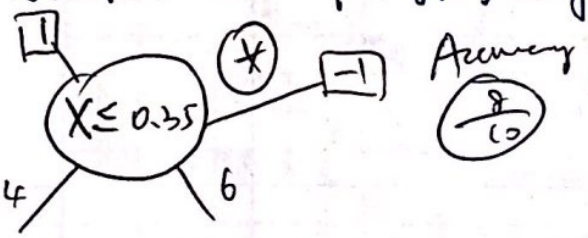
Data Mining, 2/19 Tuesday, Lecture 9

No Class on Thursday

Ensemble Methods: improve classification accuracy by aggregating the predictions of multiple classifier.

- Requirements: base classifiers better than 50% } accurate & Diverse
  All base classifiers be independent
- Types of EM: Homogeneous → ensemble classifier
  Heterogeneous → voting classifier multiple
- Bias and Variance
- Methods for constructing an Ensemble Classifier
  - By manipulating ① training set (bagging & boosting)
    ② input features (random features)
    ③ class labels (multi-class partitioning)
    ④ learning algorithm

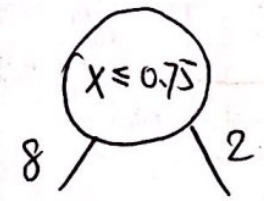
① Bagging: Bootstrap aggregating



Label 1: 4
Label -1: 0
GINI = 0

Label 1: 2
Label -1: 4
$GINI = 1 - (\frac{2}{6})^2 - (\frac{4}{6})^2$
$= 1 - \frac{1}{9} - \frac{4}{9}$
$= \frac{4}{9}$

$GINI(split) = 0 \cdot \frac{4}{10} + \frac{4}{9} \cdot \frac{6}{10}$
$= \frac{24}{90} = \frac{12}{45} = \frac{4}{15}$
$= 0.267$



① : 4
-1 : 4
$GINI = 1 - (\frac{4}{8})^2 - (\frac{4}{8})^2$
$= 1 - \frac{1}{4} - \frac{1}{4}$
$= \frac{1}{2}$

① : 2
-1 : 0
GINI = 0

$GINI(split) = \frac{8}{10} \cdot \frac{1}{2} + (\frac{2}{10}) \cdot 0 = 0.4$

Scanned with CamScanner

② Random Forests : An ~~extra~~ extension of bagged decision trees

Randomly select a subset of input features.

① → Boosting : Using weights when doing bootstrap bagging

↓

Updated weight according on previous correction

Scanned with CamScanner

Data Mining, 2/28 | Thursday, Lecture 11.

Class Imbalance Problem: Accuracy doesn't work well.

Confusion Matrix (Binary classification)

		Predicted class	
		+	−
Actual class	+	2	8
	−	1	999

Precision and Recall Trade off

Confusion

F-measure $F = 2 \times$ precision in likes

ROC Curves (Receiver ~~Operation characteristics~~

NEED

ACT

60	50
2	97

Threshold %

The sensitivity - 50/100 - 50%

TR vs. Speech speaking - 97/100 = 5%

TNR = ~~Inspector~~ ~~Specialist~~ 1 − specificity

FPR = 3/100

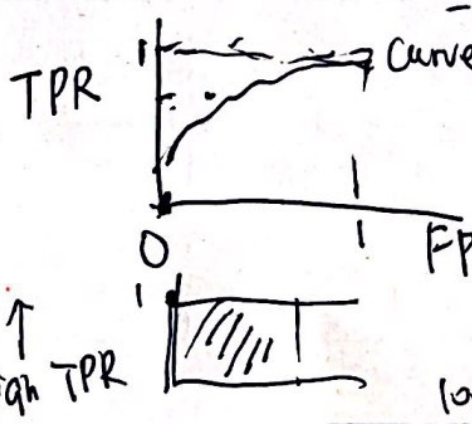
Ex: Example

	+	−	
+	2	1	TPR 2/3
−	0	3	FPR 0

Curve (1)

	+	−	
+	2	1	2/3
−	1	2	1/3

Curve (2)



Scanned with CamScanner

Mitigating Class Imbalances

- sampling based approaches.
  - * Undersampling
  - * Oversampling

→ Sampling Algorithms - SMOTE: Synthetic Minority Over-Sampling Technique.

Scanned with CamScanner