Machine Learning (ML)

October 26, 2017

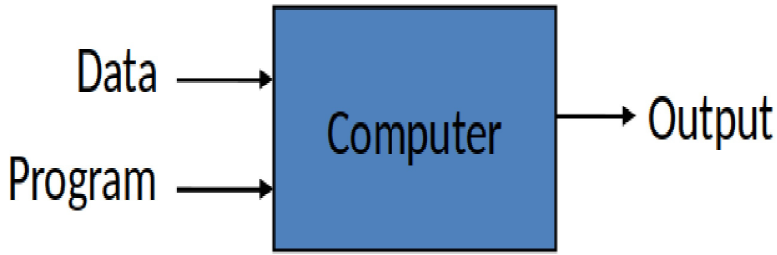
Outline

Machine Learning

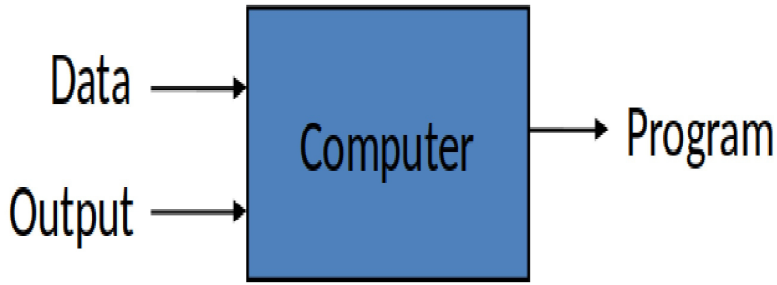
- Study of algorithms that improve their performance at some task with experience.
- Optimize a performance criterion using example data or past experience.

Traditional Programming vs ML

Traditional Programming:



ML:



Concept of learning in ML system

- Learning = Improving with experience at some task
  - Improve over task T
  - With respect to performance measure, P
  - Based on experience E.

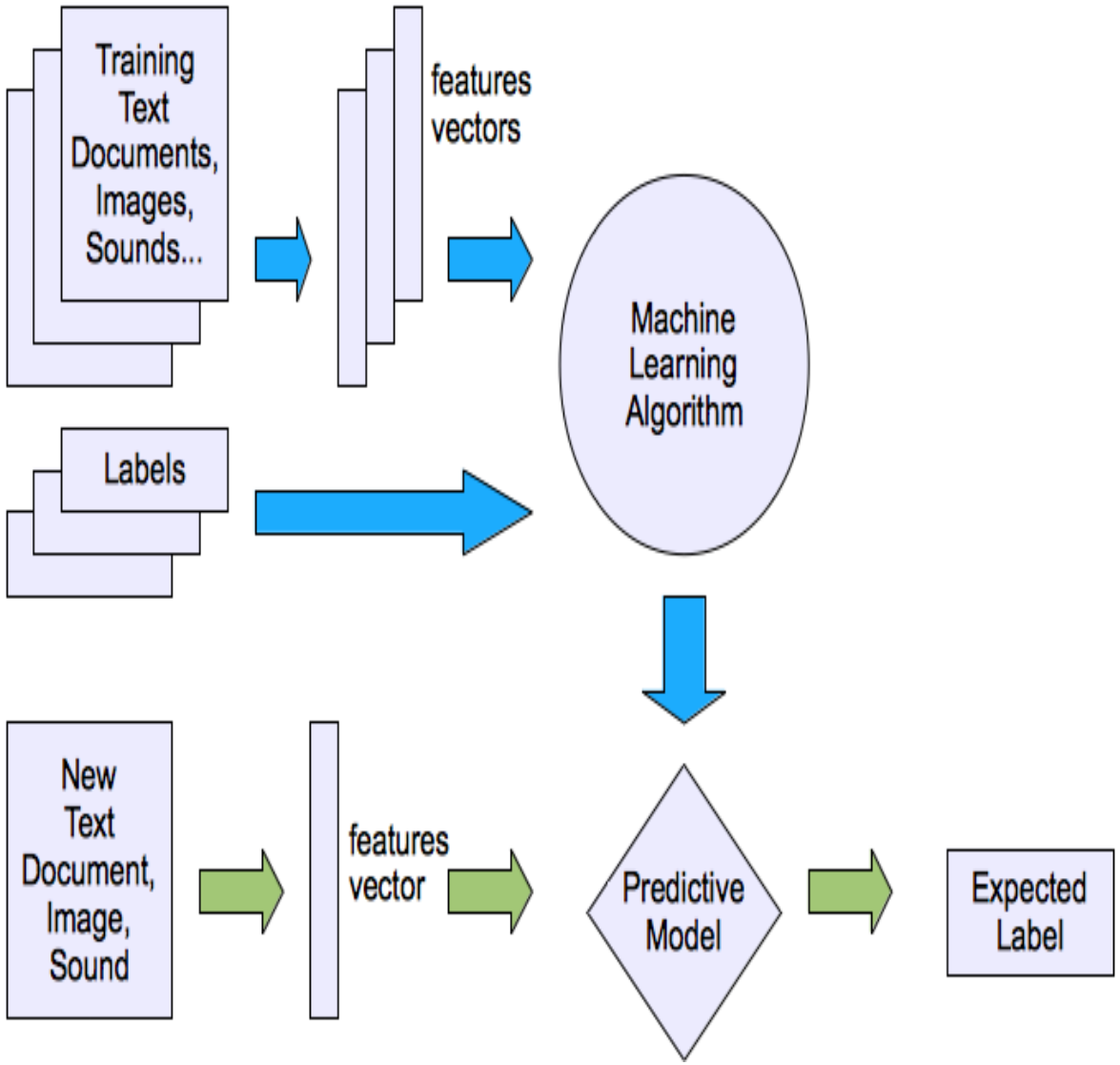
Learning to filter spam

- Spam is a mail the user does not want to receive and has not asked to receive.
  - T : Identify Spam emails
  - P : % of spam emails that are filtered correctly filtered out. Or Non spam emails are incorrectly filtered out.
  - E : a database of emails that were labelled by users.

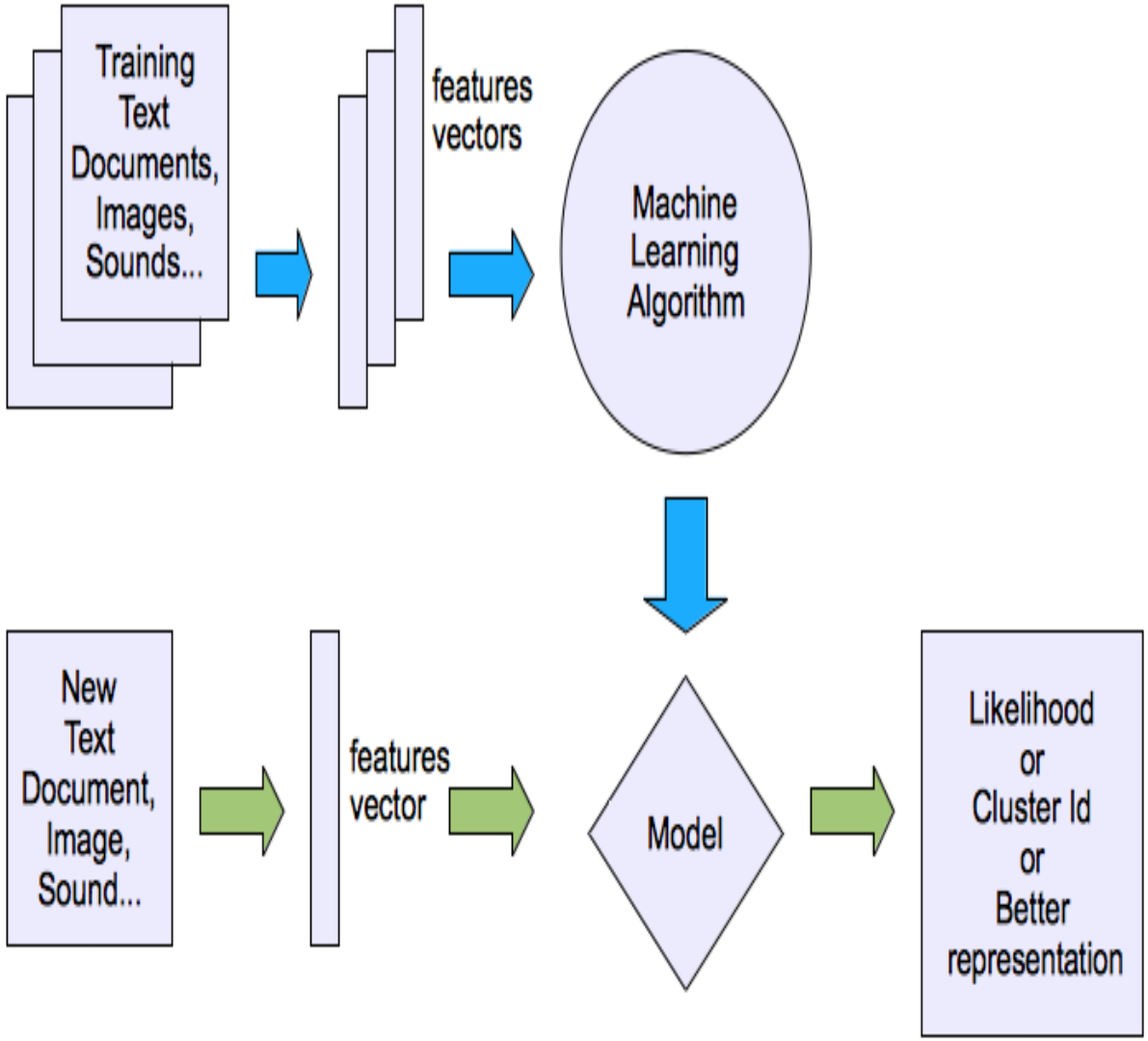
Learning Types

- Supervised Learning
  - Learning a mapping from a set of inputs to a target variable
    - Classification: target variable is discrete (e.g., spam email)
    - Regression: target variable is real-valued (e.g., stock market)
- Unsupervised Learning
  - No target variable provided
    - Clustering: grouping data into K groups
- Other types of Learning
  - Reinforcement learning: e.g., game-playing agent
  - Learning to rank, e.g., document ranking in Web search

Machine learning structure: Supervised learning



Unsupervised learning



Terminology

- Features
  - The number of features or distinct traits that can be used to describe each item in a quantitative manner.
- Feature Vector
  - N dimensional vector of numerical features that represents some object.
- Feature Extraction
  - Preparation of feature vector.
  - Transform data from higher dimensional space to fewer dimension.
- Training / Validation dataset
  - Set of data to discover potentially predictive relationships.

Representation

Student table:

R.No.	Name	Addr	M1	M2	Result
1	X	A1	68	45	F
2	Y	A2	54	34	F
3	Z	A3	76	89	P
4	P	A4	98	28	F

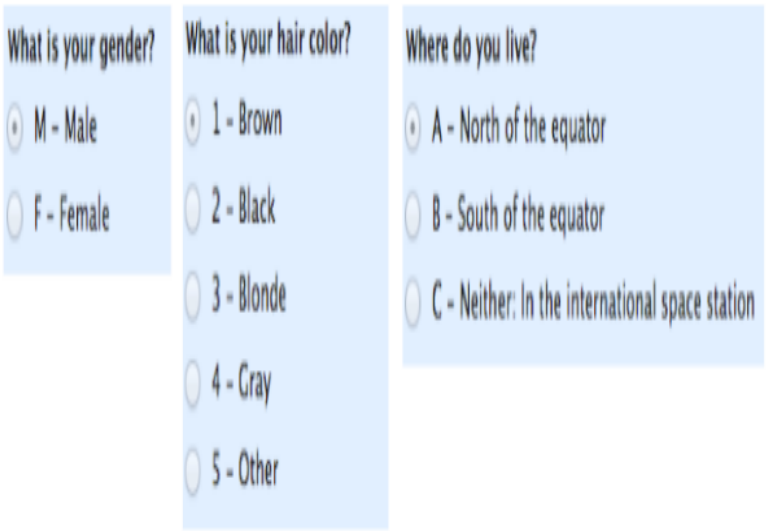
Types of Data & Measurement Scales

- Ways to categorize different types of variables.
- There are four measurement scales (or types of data).
  - Nominal
  - Ordinal
  - Interval
  - Ratio

Nominal

- Nominal scales are used for labeling variables, without any quantitative value. "Nominal" scales could simply be called "labels."
- These scales are mutually exclusive (no overlap) and none of them have any numerical significance.
- A sub-type of nominal scale with only two categories (e.g. male/female) is called "dichotomous."

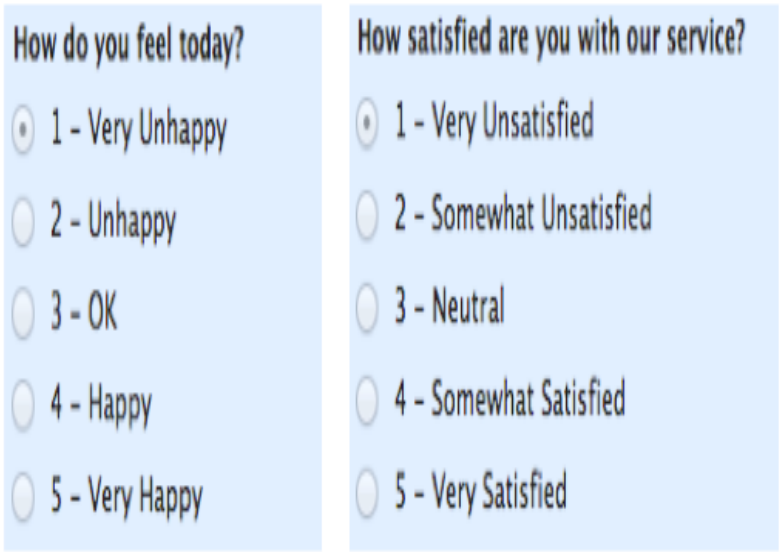
Example of Nominal Scale



Ordinal

- With ordinal scales, it is the order of the values is what's important and significant, but the differences between each one is not really known.
- In each case, we know that a #4 is better than a #3 or #2, but we don't know and cannot quantify how much better it is.
- For example, is the difference between "OK" and "Unhappy" the same as the difference between "Very Happy" and "Happy?" We can't say.
- Ordinal scales are typically measures of non-numeric concepts like satisfaction, happiness, discomfort, etc.

Example of Ordinal Scale



Interval

- Interval scales are numeric scales in which we know not only the order, but also the exact differences between the values.
- The classic example of an interval scale is Celsius temperature because the difference between each value is the same. For example, the difference between 60 and 50 degrees is a measurable 10 degrees, as is the difference between 80 and 70 degrees.
- Time is another good example of an interval scale in which the increments are known, consistent, and measurable.
- For example, central tendency can be measured by mode, median, or mean; standard deviation can also be calculated.

Example of Interval Scale

- For example, there is no such thing as "no temperature."
- Without a true zero, it is impossible to compute ratios.
- With interval data, we can add and subtract, but cannot multiply or divide.
- Consider this: 10 degrees + 10 degrees = 20 degrees. No problem there. 20 degrees is not twice as hot as 10 degrees, however, because there is no such thing as "no temperature" when it comes to the Celsius scale.

Ratio scale

- The exact value between units, and they also have an absolute zero–which allows for a wide range of both descriptive and inferential statistics to be applied.
- Ratio scales provide a wealth of possibilities when it comes to statistical analysis.
- These variables can be meaningfully added, subtracted, multiplied, divided (ratios). Central tendency can be measured by mode, median, or mean; measures of dispersion, such as standard deviation and coefficient of variation can also be calculated from ratio scales.

Comparison

Provides:	Nominal	Ordinal	Interval	Ratio
Order of value is known		Y	Y	Y
Frequency Distribution	Y	Y	Y	Y
Mode	Y	Y	Y	Y
Median		Y	Y	Y
Mean			Y	Y
Quantify difference between two value			Y	Y
Add or subtract value			Y	Y
Multiply and devide				Y
True Zero				Y

Why preprocessing

- Real world data are generally
  - Incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
  - Noisy: containing errors or outliers
  - Inconsistent: containing discrepancies in codes or names
- Tasks in data preprocessing
  - Data cleaning: fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies.
  - Data integration: using multiple databases, data cubes, or files.
  - Data transformation: normalization and aggregation.
  - Data reduction: reducing the volume but producing the same or similar analytical results.
  - Data discretization: part of data reduction, replacing numerical attributes with nominal ones.

Why Data Preprocessing is Beneficial to Data Mining?

- Less data
  - Data mining methods can learn faster.
- Higher accuracy
  - Data mining methods can generalize faster.
- Simple results
  - Easier to understand.
- Fewer attributes
  - Removing irrelevant and redundant attributes

Data cleaning

- Fill in missing values (attribute or class value):
  - Ignore the tuple: usually done when class label is missing.
  - Use the attribute mean (or majority nominal value) to fill in the missing value.
  - Use the attribute mean (or majority nominal value) for all samples belonging to the same class.
  - Predict the missing value by using a learning algorithm: consider the attribute with the missing value as a dependent (class) variable and run a learning algorithm (usually Bayes or decision tree) to predict the missing value.

Data cleaning Contd...

- Identify outliers and smooth out noisy data
  - Binning
    - Sort the attribute values and partition them into bins
    - Then smooth by bin means, bin median, or bin boundaries.
  - **Clustering**: group values in clusters and then detect and remove outliers (automatic or manual).
  - **Regression**: smooth by fitting the data into regression functions.
- Correct inconsistent data: use domain knowledge or expert decision.

Binning Methods for Data Smoothing

- Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34
- Partition into (equi-width) bins:
  - Bin 1 (4-14): 4, 8, 9
  - Bin 2(15-24): 15, 21, 21, 24
  - Bin 3(25-34): 25, 26, 28, 29, 34
- Smoothing by bin means:
  - Bin 1: 7, 7, 7
  - Bin 2: 20, 20, 20, 20
  - Bin 3: 28, 28, 28, 28, 28
- Smoothing by bin boundaries:
  - Bin 1: 4, 4, 4
  - Bin 2: 15, 24, 24, 24
  - Bin 3: 25, 25, 25, 25, 34

Binning Methods for Data Smoothing

- Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34
- Partition into (equi-depth) bins:
  - Bin 1: 4, 8, 9, 15
  - Bin 2: 21, 21, 24, 25
  - Bin 3: 26, 28, 29, 34
- Smoothing by bin means:
  - Bin 1: 9, 9, 9, 9
  - Bin 2: 23, 23, 23, 23
  - Bin 3: 29, 29, 29, 29, 29
- Smoothing by bin boundaries:
  - Bin 1: 4, 4, 4, 15
  - Bin 2: 21, 21, 25, 25
  - Bin 3: 26, 26, 26, 34

Data Integration

- Combines data from multiple sources into a coherent store
- Detecting and resolving data value conflicts
  - for the same real world entity, attribute values from different sources are different
  - possible reasons: different representations, different scales, e.g., metric vs. British units
- Redundant data may be able to be detected by correlational analysis.

Correlational Analysis

$$R_{A,B} = \frac{cov(A,B)}{\sqrt[2]{s_A^2 * s_B^2}}$$

$$Cov(A,B) = \frac{\sum(A-\overline{A})*(B-\overline{B})}{n-1}$$

$$s_A^2 = \frac{\sum(A-\overline{A})^2}{n-1}$$

$$s_B^2 = \frac{\sum(B-\overline{B})^2}{n-1}$$

Note: < 0 negatively correlated, $= 0$ no correlation, > 0 correlated – consider removal of A or B

Correlational Analysis Example

Infant ID #	Gestational Age (wks)	Birth Weight (gm)
1	34.7	1895
2	36.0	2030
3	29.3	1440
4	40.1	2835
5	35.7	3090
6	42.4	3827
7	40.3	3260
8	37.3	2690
9	40.9	3285
10	38.3	2920
11	38.5	3430
12	41.4	3657
13	39.7	3685
14	39.7	3345
15	41.1	3260
16	38.0	2680
17	38.7	2005

Correlational Analysis Example Contd...

Infant ID #	Gestational Age	$(X - \overline{X})$	$(X - \overline{X})^2$
1	34.7	-3.7	13.69
2	36.0	-2.4	5.76
3	29.3	-9.1	82.81
4	40.1	1.7	2.89
5	35.7	-2.7	7.29
6	42.4	4.0	16.00
7	40.3	1.9	3.61
8	37.3	-1.1	1.21
9	40.9	2.5	6.25
10	38.3	-0.1	0.01
11	38.5	0.1	0.01
12	41.4	3.0	9.00
13	39.7	1.3	1.69
14	39.7	1.3	1.69
15	41.1	2.7	7.29
16	38.0	-0.4	0.16
17	38.7	0.3	0.09
	$\Sigma X = 652.1$	$\Sigma (X - \overline{X}) = 0$	$\Sigma (X - \overline{X})^2 = 159.45$

Correlational Analysis Example Contd...

Infant ID #	Birth Weight	$(Y - \overline{Y})$	$(Y - \overline{Y})^2$
1	1895	-1007	1,014,049
2	2030	-872	760,384
3	1440	-1462	2,137,444
4	2835	-67	4,489
5	3090	188	35,344
6	3827	925	855,625
7	3260	358	128,164
8	2690	-212	44,944
9	3285	383	146,689
10	2920	18	324
11	3430	528	278,784
12	3657	755	570,025
13	3685	783	613,089
14	3345	443	196,249
15	3260	358	128,164
16	2680	-222	49,284
17	2005	-897	804,609
	$\Sigma Y = 49,334$	$\Sigma (Y - \overline{Y}) = 0$	$\Sigma (Y - \overline{Y})^2 = 7,767,660$

Correlational Analysis Example Contd...

Infant Identification Number	$(X - \overline{X})$	$(Y - \overline{Y})$	$(X - \overline{X})(Y - \overline{Y})$
1	-3.7	-1007	3725.9
2	-2.4	-872	2092.8
3	-9.1	-1462	13,304.2
4	1.7	-67	-113.9
5	-2.7	188	-507.6
6	4.0	925	3700.0
7	1.9	358	680.2
8	-1.1	-212	233.2
9	2.5	383	957.5
10	-0.1	18	-1.8
11	0.1	528	52.8
12	3.0	755	2265.0
13	1.3	783	1017.9
14	1.3	443	575.9
15	2.7	358	966.6
16	-0.4	-222	88.8
17	0.3	-897	-269.1
			$\Sigma (X - \overline{X})(Y - \overline{Y})$= 28,768.4

Correlational Analysis Example Contd...

$Cov(X, Y) = 1798.025$

$s_X^2 = 9.965625$

$s_Y^2 = 485478.75$

$R_{X,Y} = 0.817444625776688$

Note: Strong positive correlation.
Remove any one attribute.

Data transformation

- Normalization:
  - Scaling attribute values to fall within a specified range.
  - Scaling by using mean and standard deviation (useful when min and max are unknown or when there are outliers)
- Aggregation: moving up in the concept hierarchy on numeric attributes.
- Generalization: moving up in the concept hierarchy on nominal attributes.
- Attribute construction: replacing or adding new attributes inferred by existing attributes.

min-max normalization

$$v' = \frac{v - min_A}{max_A - min_A} * (newmax_A - newmin_A) + newmin_A$$

Example: Income, min: \$55,000, max: \$150000 -> map to 0.0 – 1.0

\$73,600 is transformed to :

$$v' = \frac{73600 - 55000}{150000 - 55000} * (1.0 – 0) + 0 = 0.196$$

z-score normalization

$$v' = \frac{v - mean(A)}{SD(A)}$$

Example: Income, mean \$33000, SD \$11000

\$73600 is transformed to :

$$v' = \frac{73600 - 33000}{11000} = 3.69$$

Data reduction

- Reducing the number of attributes
  - Data cube aggregation: applying roll-up, slice or dice operations.
  - Removing irrelevant attributes: attribute selection (filtering and wrapper methods), searching the attribute space.
  - **Principal component analysis** (numeric attributes only): searching for a lower dimensional space that can best represent the data..
- Reducing the number of attribute values
  - Binning (histograms): reducing the number of attributes by grouping them into intervals (bins).
  - Clustering: grouping values in clusters.
  - Aggregation or generalization
- Reducing the number of tuples
  - Sampling

Principal Component Analysis (PCA)

Why PCA? Independent variables are highly correlated, which affects model accuracy and reliability.

- Summarize data with many independent variables to a smaller set of derived variables.
- In such a way, that first component has maximum variance, followed by second, followed by third and so on.
- First k components explains maximum variance from total variance.
- Total variance of N independent variables = Variance of variable 1+ ... + Variance of variable N.

PCA Contd..

- Get correlation or covariance matrix
- Get eigen values
- Get eigen vectors, which are the direction of principal components
- Find coordinates of each data points in the direction of principal components.

PCA covariance matrix

$$Covaraince\ matrix A = \begin{bmatrix} Variance(X1) & Covariance(X1, X2) \\ Covariance(X1, X2) & Variance(X2) \end{bmatrix}$$

$$Variance = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$$

$$Covariance = \frac{\sum(X1 - \overline{X1}) * (X2 - \overline{X2})}{n-1}$$

PCA covariance matrix Contd...

	X1	X2	X1-Mean(X1)	X2-mean(X2)	(X1-Mean(X1)) * (X2-mean(X2))	(X1-Mean(X1)) * (X1-Mean(X1))	(X2-Mean(X2)) * (X2-mean(X2))
	1.4	1.65	1.85	2.2175	4.102375	3.4225	4.91730625
	1.6	1.975	2.05	2.5425	5.212125	4.2025	6.46430625
	-1.4	-1.775	-0.95	-1.2075	1.147125	0.9025	1.45805625
	-2	-2.525	-1.55	-1.9575	3.034125	2.4025	3.83180625
	-3	-3.95	-2.55	-3.3825	8.625375	6.5025	11.44130625
	2.4	3.075	2.85	3.6425	10.381125	8.1225	13.26780625
	1.5	2.025	1.95	2.5925	5.055375	3.8025	6.72105625
	2.3	2.75	2.75	3.3175	9.123125	7.5625	11.00580625
	-3.2	-4.05	-2.75	-3.4825	9.576875	7.5625	12.12780625
	-4.1	-4.85	-3.65	-4.2825	15.631125	13.3225	18.33980625
Mean	-0.45	-0.5675		Sum	71.88875	57.805	89.5750625

PCA covariance matrix Contd...

$$Covaraince\ matrix\ A = \begin{bmatrix} 6.42777777777778 & 7.98763888888889 \\ 7.98763888888889 & 9.95278472222222 \end{bmatrix}$$

PCA Eigen Value

$|A - \lambda I| = 0$

$(Variance(X1) - \lambda) * (Variance(X2) - \lambda) - Covarance(X1, X2)^2 = 0$

$\lambda^2 - (Variance(X1) + Variance(X2)) * \lambda + (Variance(X1) * Variance(X2) - Covarance(X1, X2)^2) = 0$

$\lambda^2 - 16.3756\lambda + 0.12214 = 0$

Solution of $a\lambda^2 + b\lambda + c = 0$

$\lambda = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$

$\lambda = \frac{16.3756 \pm \sqrt{16.3756^2 - 4 * 1 * 0.12214}}{2 * 1}$

$\lambda_1 = 16.3680, \lambda_2 = 0.007462$

Note: $\lambda_1 + \lambda_2 = 16.3756 =$ Total variance.

PCA Eigen Vector for λ_1, λ_2

For λ_1,
$(A - \lambda_1) * X = 0$

$$\begin{pmatrix} -9.9453 & 7.9876 \\ 7.9876 & -6.4153 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

We get $X_1 = 0.6262$ and $X_2 = 0.7797$

For λ_1, we get $X_1 = 0.7797$ and $X_2 = -0.6262$

Principal component

λ_1 covers 99.95% of total variance.
New Principal component is,
$[X1\ X2][\textit{Eigen vector for } \lambda_1] = [\textit{Principal component}1]$

$$\begin{pmatrix} 1.4 & 1.65 \\ 1.6 & 1.975 \\ -1.4 & -1.775 \\ -2 & -2.525 \\ -3 & -3.95 \\ 2.4 & 3.075 \\ 1.5 & 2.025 \\ 2.3 & 2.75 \\ -3.2 & -4.05 \\ -4.1 & -4.85 \end{pmatrix} \begin{pmatrix} 0.6262 \\ 0.7797 \end{pmatrix} = \begin{pmatrix} 2.163185 \\ 2.5418275 \\ -2.2606475 \\ -3.2211425 \\ -4.958415 \\ 3.9004575 \\ 2.5181925 \\ 3.584435 \\ -5.161625 \\ -6.348965 \end{pmatrix}$$

Discretization and generating concept hierarchies

- Unsupervised discretization - class variable is not used.
  - Equal-interval (equiwidth) binning: split the whole range of numbers in intervals with equal size.
  - Equal-frequency (equidepth) binning: use intervals containing equal number of values.
- Supervised discretization - uses the values of the class variable.
  - Using class boundaries. Three steps:
    - Sort values.
    - Place breakpoints between values belonging to different classes.
    - If too many intervals, merge intervals with equal or similar class distributions.
  - Entropy (information)-based discretization.
- Generating concept hierarchies: recursively applying partitioning or discretization methods.

Entropy (information)-based discretization

- Entropy based method uses a split approach.
- The entropy (or the information content) is calculated based on the class label.
- Intuitively, it finds the best split so that the bins are as pure as possible that is the majority of the values in a bin correspond to have the same class label.
- Formally, it is characterized by finding the split with the maximal information gain.

Entropy (information)-based discretization Contd...

Hours Studied	A on Test
4	N
5	Y
8	N
12	Y
15	Y

Entropy (information)-based discretization Contd...

	A on Test	**Less than A**
Overall	3	2

$$Entropy(D_1) = -\sum_{i=1}^{m} p_i log_2 p_i$$

$$Entropy(D) = -(\tfrac{3}{5} log_2(\tfrac{3}{5}) + \tfrac{2}{5} log_2(\tfrac{2}{5})) = 0.529 + 0.442 = 0.971$$

To find a split, we average two neighboring values in the list.

Entropy (information)-based discretization Contd...

Split 1: 4.5

we split at 4.5 ((5+4)/2). Now we get two bins, as follows:

	A on Test	**Less than A**
$\leqslant 4.5$	0	1
> 4.5	3	1

Entropy for each bin:

Net entropy:

$$Info_A(D) = \frac{D_1}{D} Entropy(D_1) + \frac{D_2}{D} Entropy(D_2)$$

$$Info_A(D_{new}) = \frac{1}{5} * 0 + \frac{4}{5} * 0.811 = 0.6488$$

Information gain of this split:

$$Gain_A = Entropy(D) - Info_A(D_{new})$$

$$Gain_A = 0.971 - 0.6488 = 0.322$$

Entropy (information)-based discretization Contd...

Split 2: 6.5

Average our next two values, and we get 6.5:

	A on Test	**Less than A**
$\leqslant 6.5$	1	1
> 6.5	2	1

Entropy for each bin:

$Entropy(D_{\leqslant 6.5}) = -(\frac{1}{2} log_2(\frac{1}{2}) + \frac{1}{2} log_2(\frac{1}{2})) = 1$

$Entropy(D_{>6.5}) = -(\frac{2}{3} log_2(\frac{2}{3}) + \frac{1}{3} log_2(\frac{1}{3})) = 0.389 + 0.528 = 0.917$

Net entropy:

$Info_A(D_{new}) = \frac{2}{5} * 1 + \frac{3}{5} * 0.917 = 0.9502$

Information gain of this split:

$Gain_A = 0.971 - 0.9502 = 0.0208$

This is less gain than we had before, so our best split point is still at 4.5.

Entropy (information)-based discretization Contd...

Split 3: 10

Average our next two values, and we get 10:

	A on Test	**Less than A**
$\leqslant 10$	1	2
> 10	2	0

Entropy for each bin:

$$Entropy(D_{\leqslant 10}) = -(\frac{1}{3} log_2(\frac{1}{3}) + \frac{2}{3} log_2(\frac{3}{3})) = 0.917$$

$$Entropy(D_{>10}) = -(\frac{2}{2} log_2(\frac{2}{2}) + \frac{0}{2} log_2(\frac{0}{2})) = 0$$

Net entropy:

$$Info_A(D_{new}) = \frac{2}{5} * 0 + \frac{3}{5} * 0.917 = 0.55$$

Information gain of this split:

$$Gain_A = 0.971 - 0.55 = 0.421$$

This is the clear winner at this point.

Entropy (information)-based discretization Contd...

Split 14: 13.5

Average our next two values, and we get 13.5:

	A on Test	**Less than A**
$\leqslant 13.5$	2	2
> 13.5	1	0

Entropy for each bin:

$Entropy(D_{\leqslant 13.5}) = -(\frac{2}{4} log_2(\frac{2}{4}) + \frac{2}{4} log_2(\frac{2}{4})) = 1$

$Entropy(D_{>13.5}) = -(\frac{1}{1} log_2(\frac{1}{1}) + \frac{0}{1} log_2(\frac{0}{1})) = 0$

Net entropy:

$Info_A(D_{new}) = \frac{4}{5} * 1 + \frac{1}{5} * 0 = 0.8$

Information gain of this split:

$Gain_A = 0.971 - 0.8 = 0.171$

Best split is split 3, which gives us the best information gain of 0.421.

THANK YOU