Unsupervised learning
- clustering
- density estimation

Unsupervised learning
- parametric ULearning
  1. construct gaussian mixture models
  2. use expectation maximization algo
- non-parametric

Clustering as a mixture of gaussians
"parametric (vs) non-parametric distribution".

Different statistical distributions : (next page)

*If data is ordinal/interval based, only non-parametric statistics can be used

Data
- quantitative
  - discrete
  - continuous
    - Interval scale
    - Ration scale
- qualitative
  Non-numerial
  - Nominal
  - Ordinal (order present)

Sequential learning

mapping input seq to output seq using state machines. Hidden state seq present.

Active learning.

Theory of rational agency: (action selection theories)

* Density estimation ((how using deep generative NN's?))
→ Estimating probability density function of random variable in a population from sample's help.

q: Difference between probability density fn & probability distribution fn?

q: what is maximum likelihood estimation?
→ Finding the values of parameters that result in best fit curve.

* Likelihood & loglikelihood

$$\boxed{L(\mu, \sigma; data) = P(data; \mu, \sigma)}$$

q. when is least squares minimization same as max likelihood estimation? why & how does it happen in that case?

Bayesian Inference for parameter estimation:

Bayes theorem: $P(\theta/data) = \dfrac{P(data/\theta) \times P(\theta)}{P(data)}$

* Parameter estimation using Bayes theorem
"prior distribution".
θ → set of parameters ($\theta = \{\mu, \sigma\}$ for gaussian distribution)
$P(\theta/data)$ → posterior distribution
$P(\theta)$ → prior distribution
$P(data)$ → evidence & $data = \{y_1, y_2, \dots, y_n\}$
↳ normalizing const (helps making $\Sigma P(\theta/data) = 1$)

Can we use bayesian inference for classification problems? How? Is it used for discrete data / continuous or both?

Different statistics from the posterior distribution & their physical significance?

→ expected value ≡ mean
→ variance ⇒ uncertainity
→ mode ≡ MAP estimate.

"gaussian distribution is conjugate to itself wrt gaussian likelihood function."

(Latent Dirichlet Allocation algo) ☆

Markov Chain Monte Carlo methods → to Calculate posterior distribution

☆☆ Updating beliefs iteratively in real time using bayesian inference → kalman filter.

Prior acts as a regularizer ~~here~~ in bayesian inference.

q when does MAP estimate equals MLE?

Overfitting due to Bayesian priors ⟶ $\boxed{\text{pending}}$

Marginalization

$$P(X) = \int_y P(X, Y=y)\,dy.$$

What is discriminant analysis?

When is it used?

when dependant variable is categorical & predictor/independant variable is interval.

develop discriminant fn as a linear combination of independant variables to discriminate between categories of dependant variable.

Discriminant analysis (vs) Analysis of variance. (vs) regression analysis

Correlation is not causation ✱

probability distribution

continuous

eg: Normal
Standard Normal

discrete
eg: Binomial
Poisson

MCMC methods

→ monte carlo simulations
→ markov chains (are memoryless)

* bell curve, law of large nos.

Markov → Non independant events may also conform to patterns
(In long run, dist settle to pattern)
* if events are subj to fixed prob.,
interdependant events conform to average.

q. How can bayesian inference be used to quantify uncertainity in predictions?

MCMC → Random samplig of parameters in probabilistic space to approximate the posterior distribution in bayesian inference.

~~Where~~ How can we use these posterior distributions?

→ quantifying uncertainity
→ comparing models
→ generating predictions

Central limit theorm & law of large nos

Covariance vs correlation vs causation

Statistical distributions

q: data discrete/continuous?
q. symmetry of data & outliers scenario.
q. upper & lower limits of data
q. likelihood of occurence of extreme values

Discrete distributions:
→ Binomial
→ Poisson
→ negative binomial
→ geometric
→ Bernoulli

Continuous
→ Normal
→ Exponential
→ logistic cauchy
→ gamma
→ chisquared

q. lets say we have a distribution which is generated by combining multiple commonly known distributions. How do we find those & seperate those distributions?

Joint distributions:

Discriminant analysis:

Exponential regression : $y = \alpha e^{\beta x}$ → can be converted to linear
power regression : $y = \alpha x^{\beta}$

confidence interval (vs) prediction interval (vs) tolerance interval

quantifies uncertainity of estimated skill of model

quantifies uncertainty in single forecast

q. Finding confidence interval ~~using~~ using different ~~means~~ distributions.
q. How to get confidence interval for mean given we have a sample?

Methods to get prediction intervals:
→ bootstrap resampling
→ delta method
→ bayesian method
→ mean-variance estimation method

Resampling methods:
→ cross validation → cluster based oversamplig
→ bootstraping → random under samplig
Random oversampling
SMOTE

Probability mass fn (vs) probability density fn

Hypothesis testing:
→ null hypothesis
→ p value & critical value
→ traditional testg (vs) bayesian testing

Marginal distributions
X, Y are jointly distributed random variables

Entropy → list of generalized entropies
most widely used → shannon's entropy

$$\phi_1(P) = -\sum_{i=1}^{n} P_i \log P_i$$

Decision tree → info gain → $\phi_{parent} - \sum w_i \times \phi_{child}$ $\quad w_i = \frac{n_{child}}{n_{parent}}$
→ gini index → $1 - \sum_i p_i^2$

Hinge loss → SVM

grid search → best combination of hyperparameters for best fit
→ random search

Skewness, kurtosis, coefficient of variation

Hyperparameter optimization

Data imputation methods: → for cross-sectional datasets
→ for time series data

Cases:
→ MCAR
→ MAR
→ NMAR

① Using Mean/Median values :
→ poor results on encoded categorical features

② Using Mode
→ works with categorical features
→ can introduce bias

③ Using KNN algo.
→ impyute library
KDTree
KNN sensitive to outliers unlike SVM.

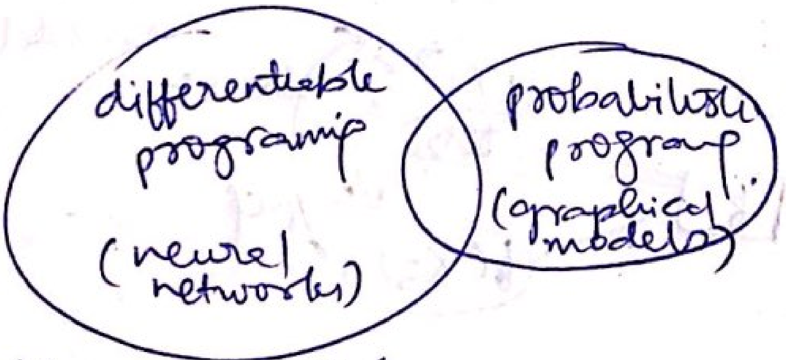
④ using multivariate imputation by <u>chained equ.</u>
→ Multiple imputations

⑤ Using DL.

⑥ using stochastic regression/extrapolation & intrapolation

<u>Uncertainity quantification</u>

→ gaussian process (GP) models.
(multivariate problems)



<u>gaussian process</u>:
→ use prior knowledg to make predictions.
→ assign probability to diff fns possible for fit a dataset, getting mean of prob dist to get most probable ans.
→ incorporate confidence.

multivariate gaussian distribution :
→ each random variable → normal distributed
→ their joint dist → also normal
→ μ → mean
Σ → covariance matrix (symmetric & +ve semidefinite)
↳ gives σ_i^2 & σ_{ij}

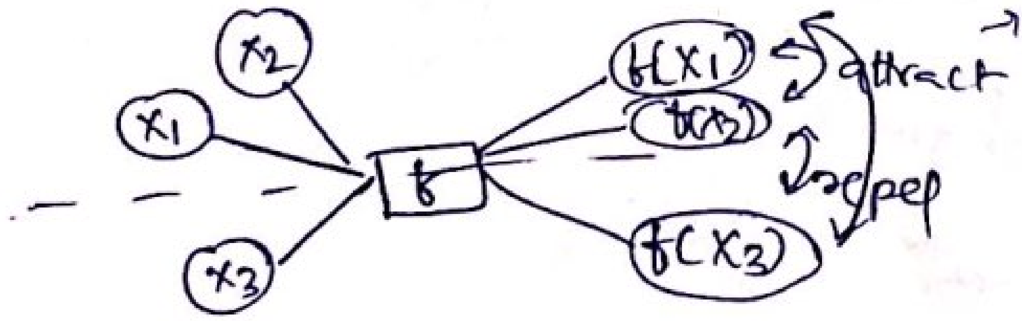
Semi supervised learning → model trained on dataset → small portion labelled data
↳ majority unlabelled data.

step 1 : cluster ~~similar~~ data into grps of similar data (unsup part)
step 2 : labelled remaining data in grp seeing the labelled data in the same grp.

manifold assumption →

contrastive learning:

adversarial ML
→ data poisoning
→ evasion attack
→ model extraction



Few short learning :
↳ meta learning problem

data level approach ← → parameter level approach
↓
→ GANs
→ data augmentation
→ using base datasets

parameter level approach
↳ limit parameter space

N-way-k-shot-Classification :
learn how to learn to classify.
Meta-learning algo

graph embedding methods & node embedding

→ learning multiple embeddings for a node

embedding
transductive ← → inductive

approaches
factorization methods ← ↓ random walks → deep learning

→ gaussian distribution based graph embedding
(includes uncertainty estimation)

for:
→ node classification
→ link prediction
→ community detection

1) Matrix factorization:
→ using adjacency matrix → most simple method
→ using (Locally linear embedding)(LLE):

$$\epsilon_i = \sum_{j \in N_i} w_{ij} \times \epsilon_j$$

$$\phi(\epsilon) = \sum_i \left(\epsilon_i - \sum_{j \in N_i} w_{ij} \times \epsilon_j\right)^2$$

2) HOPE:

$$\phi(\epsilon) = \sum_{ij} \left(\epsilon_i \epsilon_j^T - S_{ij}\right)^2$$

similarity btw nodes i, j
(using Adamic/Adar similarity)

$$A(x,y) = \sum_{u \in N(x) \cap N(y)} \frac{1}{\log|N(u)|}$$

3) DeepWalk

word embedding techniques:

- → TF-IDF
- → Word2vec → Skip-gram, → Continuous bag of words
- → GLOVE
- → BERT
- → bag of words

Gaussian process ← (set of random variables)

$$X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \sim N(\mu, \Sigma) \qquad X \sim N(\mu_X, \Sigma_{XX})$$

$$\Sigma = \mathrm{Cov}(x_i, x_j) = E[(x_i - \mu_i)(x_j - \mu_j)^T]$$

Marginalization & conditioning

$$P_{X,Y} = \begin{bmatrix} X \\ Y \end{bmatrix} \sim N(\mu, \Sigma) = N\left(\begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}, \begin{bmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{bmatrix}\right)$$

$$Y \sim N(\mu_Y, \Sigma_{YY})$$

$$P_X(x) = \int_y P_{X,Y}(x,y)\,dy = \int_y P_{X/Y}(x/y)\,P_Y(y)\,dy$$

$$X/Y \sim N\left(\mu_X + \Sigma_{XY}\Sigma_{YY}^{-1}(Y - \mu_Y),\ \Sigma_{XX} - \Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX}\right)$$

$$Y/X \sim N\left(\mu_Y + \Sigma_{YX}\Sigma_{XX}^{-1}(X - \mu_X),\ \Sigma_{YY} - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}\right)$$

each random variable has index i.

ith dimension of n dimensional multivariate distribution.

Optimization techniques

problems

single variable fns

multivariable fns

no constraints

both equality and inequality constraints

equality constraints

(lagrange multiplier method)

inequality constraints

(kuhn-tucker condns).

Numerical methods of optimization

linear program

integer program.

quadratic program

non linear program.

constraint satisfaction

infinite dim optimization

stochastic program

dynamic program.

combinatorial optimization

advanced optimization techniques

hill climbing

genetic algorithms

(inheritance, mutation, selection, crossover/recombination)

Correlation (diff from causation)

Feature selection

Supervised methods → unsupervised methods

wrapper ← → filter

Built - in feature selection

penalized regression models

lasso ← → decision tree

random forest

feature selection (vs) Dimensionality reduction

univariate statistical measures

	output numerical	output categorical
input numerical	Pearson's spearman's	anova kendall's
input categorical	Anova kendall's	chi squared mutual information

Categorical feature

nominal ← → ordinal

q. Handling categorical data in ML

nominal → one hot encoding, target encoding, binary encoding, backward difference encoding

ordinal → backward difference encoding

ordinal → label encoding

Handling missing data in time series :

→ Last observation carried forward
→ Next observation carried backward
→ Linear interpolation
→ Spline interpolation
→ when seasonality present, 1. deseasonalize 2. interpolate 3. seasonalize

Outlier analysis

univariate | multivariate

→ Normalize the data
→ check Z-score of each datapoint
→ if $x \notin [-3, 3]$, then outlier
→ using IQR
→ ORC (outlier removal clustering → uses kmeans)

Feature selection

using correlation | using P-value (not recommended)

? does it depict only linear dependency? or a more complicated function also?

Markov chain models

→ history of prev event is known.
→ prob. of transition from 1 event to another can be measured
→ prob. of future events can be computed

set of states — MC : (S, X, P) — transition prob. matrix, random variables

HMM : (N, M, A, B, μ) ← prob. for initial state

N: hidden states; M: observable event; B: prob. dist of observable output

Natural Language Understanding

- intent recognition
- entity recognition → entity
  - named entities
  - numeric entities

NLP, NLU, NLG

→ bayesian networks
→ maximum entropy
→ conditional random field
→ matrix factorization

graphical models

graph

- directed graph model (Bayesian network)
- undirected graph model (markov network)

dynamic graphs

- structure of graph changes with time
- node attributes can be time series

→ node classification → node attribut inference
→ link prediction → recommender systems travelling salesman problem
→ community detection
→ graph classification

Node classification in homogeneous graph
→ page rank, centrality measures : baseline models

manual feature engineering to augment vocabulary-based feature vectors. with graph-related node features.

* Methods to calculate quantitative values of structural position of a node in a graph.

centrality measures
examples

→ degree centrality
→ betweenness centrality
→ closeness centrality
→ eigenvector centrality

GNN → utilize ~~scalat~~ relationships in training NN on graphs (eg: GCN)

1. GCNN layer → $Z = \sigma(A'FW + b)$

node features matrix

output

activation layer

trainable parameters

graph structure (graph adjacency matrix)

Markov chain monte Carlo:

→ class of algos for systematic random sampling from high-dimensional probability distributions

→ drawing samples where next sample is dependent on prev sample.

Methods to estimate prediction interval

→ ensemble ANN

→ bayesian method] - will be looked into deeper later.

→ monte carlo method

→ bootstrap method

→ LUBE method (predicts lower & upper bounds)