

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Linear algebra MATRIX TRANSPOSITION:  * Let transpose of A is * () matrix whose (i, j) entry is * A symmetrical matrix if , * &    square matrix is symmetric  **HERMITIAN:**  Conjugate and transpose a matrix, you get:    **MATRIX MULTIPLICATION**   * Let * Pre-multiplication A to B 🡪 * Post-multiplication A to B 🡪 * :   .   * A; B are diagonal matrices of same size🡪AB = BA   REMARK:  :  **IDENTITY MATRIX:**  A identity matrix, I, is matrix whose the diagonal element is 1, otherwise is 0.  REMARK:   INVERSE OF A MATRIX:  * Scalar product: * Inverse; transpose: * Inverse of inverse:   * If A is invertible;  DETERMINANT OF SPECIAL MATRICES:   ***:***is invertible 🡪  Triangular/diagonal matrix  🡪 product of all diagonal entries  Square matrix A 🡪  Square matrix vs 2 same rows or cols  REMARK: WAYS TO DETERMINE DETERMINANT: Elementary Row Operations:    Cofactor expansion:   PROPERTIES OF DETERMINANT: , .  and | EQUIVALENT STATEMENT:  * A is invertible (A has an inverse, is singular) * There exists matrix such that * Transpose of A is an invertible matrix * has only trivial solution * The reduced row echelon form of A is * The form a basis for * The form a basis for * A has a full rank, * 0 is NOT eigenvalue of A * has a unique solution for each * The 0 * The R/C of A are linearly independent * The columns/rows of A span * The column/row space of A span * The dimension of column/row space if A is * Only vector normal to column/row space 0   ADJOINT MATRIX:  If A is invertible then  . cofactor matrix of A:   |  |  |  | | --- | --- | --- | |  |  |  |  |  |  | | --- | --- | | Cofactor matrix of A is | The adjoint matrix of A is: |  CRAMER’S RULE: ;  ; D  Answer: ; ; LU FACTORISATION: Let A be matrix.  Reduce A to row-echelon form, obtaining U matrix (should be Upper triangular matrix)    🡪  (lower triangular matrix)  Instead of solving , solve by  Step 1: solve with  Step 2: solve APPLICATION PROBLEM (DEFLECTION): We know that, by Hooke’s Law:  The results are given:   |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | |  |  |  |  |  |  |  | | Ex 1 | 1 | 0 | 1 | 0.5 | 0.3 | 0.5 | | Ex 2 | 0 | 1 | 2 | 0.1 | 0.3 | 0.7 | | Ex 3 | 2 | 1 | 0 | 0.7 | 0.3 | 0.1 |   Question: FIND D  ***Step 1***: 🡪 ( )  ***Step 2:*** Find a, b, c such that  . For example:    ***Step 4:*** combine step 2& 3, we will have: | LINEAR SPAN: be a set of vectors in set containing all linear combination of is **linear span**  of S/linear span of  To check if , check whether augmented matrix formed by S; any is inconsistent (last column is a pivot column)  ;    Matrix M:  🡪 prove  , then prove that each vector in linear combination of vectors in SUBSET; SPAN: There are 2 main representation of a subset:  Implicit:  Explicit:  Want to prove that subset = span (S) then we transfer into explicit way to compare SUBSPACE: Let V be subset of . If there exists a set of vectors such that then V is said to be a subspace of  V; W is subspaces of  🡪 is subspace of  🡪 is subspace of  🡪 is NOT subspace of | CHECK WHETHER V IS A SUBSPACE OR NOT? Check V contains vector 0  Check combination of vectors in V ABSTRACT DEFINITION OF SUBSPACE: Let V be non-empty subset of . Then V is said to be a subspace of if; only if for pair vector . LINEAR INDEPENDENCE: . S is called linearly independent set if is ONLY answer.  A is invertible 🡪 are linearly independent REDUNDANCY: are vector taken from . If is a linear combination of is redundantCheck Whether Vector Set Is Independent or Not? Ex  ( are linearly independent)  ***Step 1:*** we put vector in S in homogenous equation:  ***Step 2:***  🡪 This equation only has 1 trivial solution:  (u, v, w is independent)  NOTE: linear equation has only trivial solution independent. If not 🡪 dependent  BASIS:.  Let S = be subset in vector space V. S is called a basis S is linearly independent, S spans V.  Basis for a vector space = smallest possible # vectors that can span V  If U; W: subspaces in , there exists a basis such that is a basis for , is a basis for COORDINATE VECTORS: Let S = basis (V) &    The coefficient Vector is co-ordinate vector of v relative to basis S.  If S is a basis for then every vector has a unique co-ordinate vectors relative to V  V can have different basis 🡪 different co-ordinate vectors EUCLIDIAN VECTORS:  DIMENSION: Let V be a vector space having a basis with k vectors.  🡪 linearly dependent  🡪 cannot span V  The **dimension** of vector space = # vectors in basis of      ROW SPACE; COLUMN SPACE: FIND A BASIS FOR A ROW SPACE: Find REDUCED row-echelon form is B, then basis of row space of B = basis row space of A COLUMN SPACE: Let & be columns of A; B respectively. With each belongs to column space of AFIND A BASIS OF COLUMN SPACE OF A: Find REDCUED row-echelon form of A is B, then choose pivot column of B & take corresponding column(A) RANK OF A MATRIX:  * , * matrix, * matrix, 🡪 full rank | COLUMN SPACE; LINEAR SYSTEM: **Theorem:** a system of linear equations is consistent if; only if b lies in column space of A, or A; augmented matrix has same rank COLUMN SPACE; LINEAR SYSTEM: We have equation :  C**onsistent** 🡪b is linear combination of cols(A)  The linear system is **inconsistent** SPAN-DIMENSION-LINEAR INDEPENDENCE: If there exists a set that spans V, then  If there exists a linearly independent set in V, then  If . then there exists a set p + 1 vectors in V that spans V. NULL-SPACE OF A MATRIX: Let A be matrix. be a homogenous linear system   * The solution set of is a subspace of ; also, solution space of , **null space (A)** * . * = # pivot columns of A, nullity (A) = # non-pivot columns of A     RANK VS. NULLITY:         **SOLUTION FOR WITH NULL-SPACE:**  Let be general solution for, let be solution to equation: , general solution to is  **EUCLIDIAN NORM DISTANCE:** of vector  **DOT PRODUCT AND PROJECTION:**  The dot product of two P-dimensional vectors  is  The Euclidian norm of a vector can be computed using dot product, as ORTHOGONALITY: S = orthogonal basis, :   ORTHOGONAL PROJECTION: Let V be a subspace of then if is an orthogonal basis for V, then projection of w onto V is p: FIND DISTANCE FROM A POINT TO A LINE/PLANE: *For a line,* we only have 1 basis:  ***Step 1:*** projection  ***Step 2:*** compute distance  *For a plane,* we have 2 bases, however, alternative method: we can use “already known” orthogonal vector:  ***Step 1:*** projection w onto n,  ***Step 2:*** take length of p: | ORTHOGONAL/ORTHONORMAL SET:  * A set of vectors are mutually **orthogonal** is every pair of vectors is orthogonal * Orthogonal/orthonormal set is a basis for because:   They are set of non-zero vectors  Linearly independent set MATRIX WITH ORTHONORMAL COLS: (A **don’t** have to be square matrix) has orthonormal vectors columns if its Gram matrix is I ORTHOGONAL MATRICES: has orthonormal columns   1. Properties: is orthogonal matrix  * is also orthogonal matrix * If is an eigenvalue of Q, then * are orthogonal matrix, * **Rows of A** are an orthogonal matrix   NOTE:has orthonormal cols is NOT **orthogonal matrix**  *Product of orthogonal matrices:* If are **orthogonal matrices**; of equal size, then product: is **orthogonal matrix**  *Linear equation with orthogonal matrix:* BEST APPROXIMATION: is best approximation of  : p is projection ORTHOGONAL basis (GRAM-SCHMIDT process): span{u}  NOTE: *if you are asked to find* ***orthonormal*** *then have to make each vector in* ***orthogonal*** *become* ***unit vector***  **MATRIX-VECTOR PRODUCT**:              A has a linearly independents columns: ORTHONORMAL VECTORS:  * The vectors have unit norm * Mutually orthogonal: * **For ex,** a set of orthogonal 🡪 span vectors orthogonal tospan  ORTHOGONAL complements & ORTHOGONAL projections Let W be a subspace of Rn. **set of all vectors** that are **orthogonal to W** is called **orthogonal complement of W,** denoted . That is,  *Properties:* Let W be a subspace of   * is a subspace of * & * Let A: matrix. Then orthogonal complement of row space of A is null space of A; orthogonal complement of column space of A is null space of : |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| LEAST SQUARE SOLUTION:  * is **inconsistent**  b is not in column space of Aproject b onto column space of A solve * is least square : * **Theorem:** let be a linear system. Then x is a least square solution to if; only if x is a solution to * If linear system is **consistent** then solution set of is **same** as  QR FACTORIZATION: ***Step 1:*** We have A as a matrix, treat each column as a vector  ***Step 2:*** find **orthonormal basis** (unit vector of orthogonal basis)  ***Step 3:*** Write each of as a linear combination of . They should be in form like this:  , .  .  ***Step 4:*** write A = QR (Q is orthogonal matrix), R is triangular matrix, **+ entries** QR factorisation + LEAST SQUARE solution: NOTE: Q: orthogonal matrix 🡪  We have 🡪 compute solution for Least Square solution: EIGENVALUE & EIGENVECTOR:  * is called **eigenvalue**u is **eigenvector** associated with **eigenvalue** * The **eigenvalue** of a **diagonal matrix** are all entries on main diagonal   is an eigenvalue of  is eigenvalue of   * matrix has eigenvalues , eigenvectors If ; ,  FIND AN EIGENVALUE OF A MATRIX:  * Eigenvalue of A: that makes matrix () singular 🡪 * : **Characteristic polynomial** A * : **Characteristic equation**  EIGENSPACE:  * : singular 🡪 system has infinite solution. * The solution space : **eigen-space** of A associated with eigenvalue () 🡪 = ALL eigenvectors of A * is called a null-space of * For each we find  :  * P , D is a diagonal matrix with entries are * A is diagonalizable if; only if A has linearly independent eigenvectors /n distinct eigenvalues  CHECK WHETHER A IS DIAGONALIZABLE? ***Step 1:*** Find eigenvalues   * (order of matrix A)   ***Step 2:*** For each we find a basis for eigenspace   * : **algebraic multiplicity** of eigenvalue * : **geometric multiplicity** of * Then for each   ***Step 3:*** Let , is number of vectors in S.   * If 🡪 A is NOT diagonalizable * If 🡪 A is diagonalizable.   Equality happens:    **geometric multiplicity = algebraic multiplicity** for all eigenvalues of A If 🡪 NEVER HAPPEN | ORTHOGONAL SYMMETRIC matrix:  * A square matrix A is orthogonally diagonalizable orthogonal matrix Q such that is a diagonal matrix * **If a matrix A is orthogonally diagonalizable, then A is symmetric** (NOTE: Q is orthogonal matrix).   A is symmetric.  Eigenvectors of a symmetric matrix corresponding to different eigenvalues are orthogonal. METHOD for ORTHOGONAL diagonalization of a symmetric matrix.  1. Find eigenvalues of A. 2. Find eigenspace for each eigenvalue. 3. For repeated eigenvalues   (when dimension of eigenspace > 1)  🡪 apply Gram–Schmidt orthogonalization to find an orthogonal basis.   1. These orthogonal bases of eigenspaces form an orthogonal basis of . 2. Normalize, dividing each vector of basis by its length. 3. We have: , where D is diagonal with eigenvalues of A  REAL LIFE PROBLEM:  * D is a diagonal matrix; diagonal entries of D is   Then we have: . D is diagonal matrix so (APPLICATION)TILES:  * I have 3 kinds of tiles: red-colored tiles (1R), blue-colored tiles (2B); green-colored tiles (2G). * Let = # different ways to tile a pavement. For example, * = 1; = 3   Find .  ANSWER  ***Step 1:*** find relation between  We have already known ways to tile pavement. Now if we want to tile then: we can add on 1R, or take out 1R from make it ; add 2G/2B. since they are 3 ways (1 for ; 2 for ) then total ways is:  ***Step 2:*** set up  .  ***Step 3:*** set up :  Find eigenvalue:  P?  *.*  Find should be  Set up    .  Alternative method:    When :     Problem 1: rate of change:CONIC EQUATION: We have . general form of conic section: | Linear differential equation:  * Differential equation: * If is an egienvalue of A, associated withis vectorspace then 🡪 general solution is: () are vectors. Then we use initial condition to find (  FUNDAMENTAL SET: *Square matrix A:*   * n linearly independent functions in fundamental * S is an n-dimensional vector space of functions * If vector is specified, initial value problem is to construct a unique Y such that  PROPERTIES OF COMPLEX VECTORS:  * &  REVIEW ABOUT COMPLEX NUMBER: The polar form of a complex number:    Complex exponential form:   COMPLEX EIGENVALUE:  * If is an eigenvalue of A; is eigenvector associated with , then is an eigenvalue of A; is eigenvector associated with * Furthermore, we all know that are both conjugate solutions of linear combination of = a solution to this equation (if we don’t have initial condition) * Consider following linear combination of :  Application (Complex Eigenvalue): .          .  The general solution is . When t = 0 then: PROBLEM 2: : ***Step 1:*** Let .  ***Step 2:*** We have: .  ***Step 3:*** Find another equation relating to , we have  Then solve as usual.  **EXERCISES:** IDEMPOTENT MATRICES: A matrix A is said to be idempotent when:  A idempotent is idempotent  A is idempotent 🡪 is invertible SUBSPACES: Let V; W: subspaces of . Define:  Show that is a subspace of | Show that  Use abstract definition of subspace  ***Step 1:*** Show that , true because 0 always belong in  ***Step 2:*** Let u; v be any two vectors in ; let a; b be any real numbers.  u; v V, V . Similarly, W. Thus .  By abstract definition of subspaces, is a subspace of  If V; W: subspaces, basis for V; basis for W such that is basis & is basis  be a basis for . By adding in vectors successively, there exists vectors | is a basis for V; there exists vectors | is a basis for V.    Consider vector equation  are linearly independent, basis of  Let independent vectors in .  If A is invertible matrix are linearly independent (1) RANK, NULLITY: be cols of A; B. Show:    Show that            Show that       NULLSPACE: Show of A =  Let u: vector of of A,  Let v be vector of        nullspace of is subspace of of A  Explain why every vector in of B is also in of AB. Is this also true for every vector in of A  Suppose a vector x is in of B, then we get . By matrix multiplication, x is also in of AB.  This is NOT true for every vector in of A. Take . Then , but x is not in of AB. | *Let A; B be matricies.*  1. Show that AB = 0 if; only if column space of B is a subspace of of A   For this part of problem let columns of B be . Then ]. ⇔ *All elements of column space of B must be contained in of A.*   1. Show that if AB = 0, then sum of ranks of A; B cannot exceed n.                 Let A be non-singular matrix; B be an matrix. Prove that AB; B have same null space.  ***Step 1:*** Null space(B) is a subset of null space (AB).  If v is in null space of B, then ; hence, . Thus, v is also in null space of AB.  ***Step 2:*** Null space (AB) contained in null space(B)  If v is in . matrix A is non-singular; = 0. It follows that = 0; hence v is also in null space of B EIGNEVALUES; IDENTITY MATRIX A is a diagonalizable  matrix; has only 1; −1 as eigenvalues. *Show that .*  Since A diagonalizable has eignevals   THE EXPONENTIAL OF A MATRIX: Compute  ***Step 1:***  matrix  ***Step 2:***  ***Step 3:*** convert into exponential form:  ***Step 4:*** using matrix multiplication to calculate: RECURRENCE OF DETERMINANT: .  Let be Using cofactor expansion, thus:  **ORDINARY LEAST SQUARES:**  Linear regression models output, or target variable as a linear combination of - input . Let X be matric with each row an input vector (with a one in first position), and similarly let be vector of outputs in training set, linear model will predict given X using parameter vector, or weight vector according to  Where are residuals, or errors of prediction. is found by minimizing an objective function, which is loss function, error measured on data. This error is sum of squared errors SSE loss. Minimizing SSE is Ordinary Least Square OLS regression as objective function.  Which is a simple ordinary least squares OLS minimization. |  |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Machine Learning Introduction  **METRIC, MEASUREMENT, MEASURE:**  METRIC: unit of measurement providing way to objectively quantify performance  MEASUREMENT: act of obtain data associated vs metric  MEASURES: numerical value associated with metric.  **DATA CLASSIFICATION by measurement scales:**  CATEGORICAL (NOMINAL) DATA - sorted into categories according to specified characteristics.  ORDINAL DATA - can be ordered or ranked according to some relationship to 1 another.  INTERVAL DATA - ordinal but have constant differences between observations; have arbitrary zero points.  RATIO DATA - continuous; have natural zero.  **DATA RELIABILITY; VALIDITY:**   |  |  |  |  | | --- | --- | --- | --- | | (1st) RELIABILITY: Data is accurate; consistent  (2nd) VALIDITY: Data correctly measures what it is supposed to measure. | | | | | A tire pressure gage that consistently reads several pounds of pressure below true value |  | ✓ | | Number of calls to customer service desk (counted correctly) used to assess customer dissatisfaction | ✓ |  | | Customer rating on food quality is used to assess customer satisfaction |  |  |   **DESCRIPTIVE ANALYSIS:** characterise, consolidate; classify data to convert it into useful information for purposes of understanding; analysing business performance.  **Measures of Location** (Mean, Median, Mode)  *Symmetrical*, *unimodal*,  *Negatively* *skewed (left skewed, tails off toward right)*,  *Positive* *skewed (right skewed, tails off toward left)*,  **Measures of Dispersion** (Range, Variance, Standard deviation, Theorem, Coefficient of Variation)  **Measures of Shape** (Skewness, Kurtosis)  **Measures of Association** (.; Correlation)  **DESCRIPTIVE statistics for categorical data:**  PROPORTION is fraction of data that have certain characteristic, are key descriptive statistics for categorical data, i.e. defects or errors in quality control applications or consumer preferences in market research.  **PREDICTIVE ANALYSIS:**  ⎯Seeks to predict future by examining historical data, detecting patterns or relationships in these data; then extrapolating relationships forward in time  ⎯Predictive analysis can predict risks; find relationships in data not readily apparent with traditional analysis  ⎯Using advantaged techniques, predictive analysis can help detect hidden pattern in large quantities of data to segment; group data into coherent sets to predict behaviour; detect trends  **PRESCRIPTIVE ANALYSIS**: optimise model (minimise expenditure, maximise benefit/profit)  **MACHINE LEARNING:** involves building mathematical models to help understand data. “Learning” enters fray when we give these models that can be adapted to observed data; in this way program can be considered to be “learning” from data. Once these models have been fit to previously seen data, they can be used to predict and understand aspects of newly observed data. | **TYPES OF LEARNING:**  There are 3 types of learnings:  **SUPERVISED learning**: Models that can predict labels based on labelled training data  ⎯ set of data (training data) consists of a set of input data and correct responses corresponding to every piece of data.  ⎯ Based on this training data, algorithm must generalize such that it is able to correctly (or with a low margin of error) respond to all possible inputs.  ⎯ In essence: algorithm should produce sensible outputs for inputs that weren't encountered during training. ⎯ Also called learning from exemplars  ⎯ It consists of  **UNSUPERVISED learning**: modelling features of a dataset without reference to any label and is often described as “letting dataset speak for itself.” These models include tasks such as *clustering* and *dimensionality reduction.* Clustering algorithms identify distinct groups of data, while dimensionality reduction algorithms search for more succinct representations of data  ⎯Conceptually Different Problem.  ⎯ No information about correct outputs are available.  ⎯ No Regression No guesses about function can be made ⎯No information about correct classes. But if we design our algorithm so that it exploits similarities between inputs so as to cluster inputs that are similar together, this might perform classification automatically ⎯In essence: aim of unsupervised learning is to find clusters of similar inputs in data without being explicitly told that some data points belong to one class and other in other classes. algorithm must discover this similarity by itself  ⎯ It consists of  **SEMI-SUPERVISED learning (REINFORCEMENT learning)**: fall somewhere between supervised learning and unsupervised learning. Semi-supervised learning methods are often useful when only incomplete labels are available.  ⎯Stands in middle ground between supervised and unsupervised learning.  ⎯The algorithm is provided information about whether answer is correct but not how to improve it  ⎯The reinforcement learner must try out different strategies and see which works best  ⎯In essence: algorithm searches over state space of possible inputs and outputs in order to maximize a reward  **CLASSIFICATION task: predicting DISCRETE label**  (Supervised learning): Models that predict labels as two or more discrete categories  Some important classification algorithms: Gaussian naive Bayes (Naive Bayes Classification), support vector machines (Support Vector Machines), and random forest classification (Decision Trees and Random Forests)  **Classification problems**  🟑Consists of taking input vectors and deciding which of N classes they belong to, based on training from exemplars of each class.  🟑 Discrete (most of time). i.e. example belongs to precisely one class and set of classes covers whole possible output space.  🟑How it's done: Find 'decision boundaries' that can be used to separate out different classes.  🟑Given features that are used as inputs to classifier, we need to identify some values of those features that will enable us to decide which class current input belongs to | **REGRESSION task: predicting CONTINUOUS label**  (Supervised learning); Models that predict continuous labels  Some important regression algorithms: linear regression, support vector machines (Support Vector Machine) and random forest regression (Decision trees and random forests)  Regression problems:  🟑Given some data, you assume that those values come from some sort of function and try to find out what function is.  🟑In essence: You try to fit a mathematical function that describes a curve, such that curve passes as close as possible to all data points. So, regression is essentially a problem of function approximation or interpolation  **CLUSTERING task: Inferring labels on unlabelled data:** (Unsupervised learning)  Classification and Clustering involves building models that will predict labels for new data. 🡪 Supervised learning.  Unsupervised learning will build models without reference to any known labels 🡪 Clustering.  Some important clustering algorithms: K-means clustering, Gaussian Mixture  **DIMENSIONALITY reduction: Inferring structure of unlabelled data**  Models that detect and identify lower-dimensional structure in higher dimensional data.  **NEURAL NETWORK:**  **Motivation:** Animals learn and learning occurs within the brain If we can understand how the brain works then there are probably things that we can copy and use for our machine learning system.  The brain is massively complex and impressively powerful, But the basic atomic building blocks are simple and easy to understand.  The brain does exactly what we want it to. It deals with noisy and inconsistent data, and produces answers that are usually correct from very high dimensional data (like images) very quickly  **Basic Processing unit of the brain are neurons**  Each neuron can be thought of as a processor. Each performing a very simple computation: deciding whether to fire or not.  The brain is hence a massively parallel computer made up of billions of 'processors'  **How does learning occur in the brain?**  Plasticity: modifying the strength of connections between neurons and creating new connections  **Rule:** “Changes in the strength of interneuron (synaptic) connections are proportional to the correlation in the firing of the two connecting neurons.”  Basically: “Neurons that fire together, wire together” | Supervised learning  **SUPERVISED MACHINE LEARNING MODELS:**  Step 1: Choose a class of model  Step 2: Choose model hyperparameters.  Step 3: Fit model to training data  Step 4: Use model to predict labels for new data  The first two pieces of this—the choice of model and choice of hyperparameters—are perhaps most important part of using these tools and techniques effectively. In order to make an informed choice, we need a way to *validate* that our model and our hyperparameters are a good fit to data. While this may sound simple, there are some pitfalls that you must avoid doing this effectively.  **VALIDATION vs CROSS-VALIDATION:**  When we use validation process, we split data into 2 portions, one is for ‘Training” and one for “Testing”, we call this hold set, which is similar to unknown data, as model has not “seen” it before.  One disadvantage of using a holdout set for model validation is that we have lost a portion of our data to model training. In previous case, half dataset does not contribute to training of model! This is not optimal and can cause problems—especially if initial set of training data is small.  One way to address this is to use *cross-validation*—that is, to do a sequence of fits where each subset of data is used both as a training set and as a validation set    What comes out are two accuracy scores, which we could combine (by, say, taking mean) to get a better measure of global model performance. This form of cross-validation is a *two-fold cross-validation*—one in which we have split data into two sets and used each in turn as a validation set. We could expand on this idea to use even more trials, and more folds in data    Here we split data into five groups, and use each of them in turn to evaluate model fit on other 4/5 of data  Repeating validation across different subsets of data gives us an even better idea of performance of algorithm.  **SELECTING BEST MODEL:**  Of core importance is following question: *if our estimator is underperforming, how should we move forward?* There are several possible answers:  • Use a more complicated/more flexible model  • Use a less complicated/less flexible model  • Gather more training samples  • Gather more data to add features to each sample  answer to this question is often counterintuitive. Sometimes using a more complicated model will give worse results and adding more training samples may not improve your results! ability to determine what steps will improve your model is what separates successful machine learning practitioners from unsuccessful. | Example: We see below 2 models.  The first model attempts to find a straight-line fit through data. Because data are intrinsically more complicated than a straight line, straight-line model will never be able to describe this dataset well. Such a model is said to ***underfit* data**; 🡪 it does not have enough model flexibility to suitably account for all features in data 🡪 **high *bias***    The second model attempts to fit a high-order polynomial through data.  Here model fit has enough flexibility to nearly perfectly account for fine features in data, but even though it very accurately describes training data, its precise form seems to be more reflective of particular noise properties of data rather than intrinsic properties of whatever process generated that data. Such a model is said to ***overfit* data**; 🡪 has so much model flexibility that model ends up accounting for random errors as well as underlying data distribution 🡪 **high *variance***.    The score here is  **score**, of coefficient of determination, which measures how well a model performs relative to a simple mean of target values.  indicated perfect match  indicates model does no better than simply taking mean of data, and negative values mean even worse models. From scores associated with these two models, we can see that holds more generally:  For high-bias model, performance of model on validation set is like performance on training set.  For high-variance model, performance of model on validation set is far worse than performance on training set.  **VALIDATION CURVE:**  The expected training score and validation score is shown below:    The training score is everywhere higher than validation score this is generally case: model will be better fit to data it has seen than to data it has not seen.  For very low model complexity (a high-bias model), training data is underfit, which means that model is a poor predictor both for training data and for any previously unseen data. | For very high model complexity (a high-variance model), training data is overfit, which means that model predicts training data very well but fails for any previously unseen data.  For some intermediate value, validation curve has a maximum. This level of complexity indicates a suitable trade-off between bias and variance.  Training scores then model complexity  The validation scores reach a maximum before dropping off as model becomes overfit.  However, validation scores’ behaviour depends on two factors: model complexity and number of training points.  **LEARNING CURVES:**  The expected training score and validation score is shown below:    The general behaviour we would expect from a learning curve is this:  A model of a given complexity will overfit a small dataset: this means training score will be relatively high, while validation score will be relatively low.  A model of a given complexity will underfit a large dataset: this means that training score will decrease, but validation score will increase.  A model will never, except by chance, give a better score to validation set than training set: this means curves should keep getting closer together but never cross.  When learning curve has already converged (i.e., when training and validation curves are already close to each other), *adding more training data will not significantly improve fit!* |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Probabilities  **PERMUTATION:** A permutation of set of objects is ordering of objects in row.  of set of elements are:  REMARK:  **COMBINATION:** of set of n elements  **BINOMIAL coefficient:**  **NUMBER OF ELEMENTS IN POWER SET:** , if set S has n elements, total # subset of S  REMARK:  **NUMBER OF INTEGER SOLUTIONS: #** non-negative integer solutions of equation OR # r-combinations with repetition allowed that can be selected from a set of n objects  **ARRANGING IN A CIRCLE:** For n distinct objects arranged in a circle, there are  **CONDITIONAL PROBABILITY** of B given that A is  **GENERAL MULTIPLICATION RULE:**    **INVERSE PROB:**  **INDEPENDENT vs. MUTUALLY EXCLUSIVE:**  Being *independent*; *mutually exclusive* are *NOT same thing*.      If A & B are *mutually exclusive & non-trivial (positive prob)* then A & B *cannot*be *independent*.  **PAIRWISE INDEPENDENT EVENTS:**  A set of events are said to be pairwise independent  **MUTUALLY INDEPENDENT EVENTS:**  A set of events are said to be *mutually independent/ independent*  are mutually independent  different cases.  Mutually independence pair-wise independence  Pair-wise independence mutually independence  **PARTITION:** If are *mutually**exclusive* ; *exhaustive*  a *partition* of S.  **RULE OF TOTAL PROB.:** If is *partition*  **BAYES’S THEOREM:** Let be partition of S.  **CHEBYSHEV’S THEOREM**: Proportion of values that lie within standard deviations of mean are at least  Why is this useful? Able to use mean; standard deviation to find percentage of total observations that fall within given interval about mean | Univariate statistics  PRE-REQUITSIVE: STATISTICS & prob.  *Discover associations between a variable of interest and potential predictors. It is strongly recommended to start with simple univariate methods before moving to complex multivariate predictors.*  *Most of univariate statistics are based on linear model which is one of main model in machine learning*  **RANDOM VECTORS & RANGE SPACE**  Let be an experiment; a sample space. a two-dimensional random vector. range space is  **INDEPENDENT RANDOM VARIABLE:**  X; Y are independent  X; Y are independent  are independent    **PERCENTILES:** *kth percentile is value at or below which at least k percent of observations lie.*  COMPUTING PERCENTILES:  Find kth percentile for variable in sample size n  Rank of  **:**  *quantile of random variable X is :*  BREAK DATA INTO 4 PARTS  25th percentile, Q1; 50th percentile, Q2;  75th percentile, Q3; 100th percentile, Q4.  **VARIANCE** ~ average of squared deviations from mean. If sample data is also population data, then n = N to compute population variance  **STANDARD DEVIATION** ~ square root of variance (popular measure of risk)  **STANDARD ERROR:**  **STANDARDIZED VALUES, Z-SCORE**, provides relative measure of distance observation is from mean (independent of units of measurement)  **COEFFICIENT VARIATION** provides relative measure of dispersion in data relative *to mean:*  Provides relative measure of risk to return  Useful when comparing variability of two or more data sets with different scales  Smaller CV 🡪 smaller risk  Reciprocal of CV 🡪 return to risk  **COVARIANCE** is measure of linear association between two variables, X; Y.  POSITIVE covariance 🡪 direct relationship  NEGATIVE covariance 🡪 inverse relationship  Magnitude 🡪 degree of association            **independent**    **CORRELATION** is measure of linear association between two variables, X; Y (not dependent on units of measurement)  RANGE: (Strong negative); 1 (Strong positive linear relationship); 0 indicates no linear relationship;  Also known as: Pearson product moment correlation or Pearson's correlation coefficient  If X; Y are **independent**, then .  *On other hand, does not imply independence.* | **DISPERSION:** Dispersion refers to degree of variation (numerical spread/ compactness)  RANGE data values  INTERQUARTILE RANGE IQR difference between first; third quartiles: (use 50% data)  **SKEWNESS:** describes lack of symmetry of data.  COEFFICIENT OF SKEWNESS (CS):  left-skewed data; right-skewed  suggests high degree of skewness.  suggests moderate skewness.  suggests relative symmetry  **KURTOSIS:** refers to (high, narrow, or flatness)/(short, flat-top) of histogram  **COEFFICIENT OF KURTOSIS CK:** measures degree of kurtosis of population  CK <3 🡪 data is flat and wide degree of dispersion  CK >3 🡪 data is peaked with less dispersion  **OUTLIERS:** Mean; Range are sensitive to outliers  HOW DO WE IDENTIFY POTENTIAL OUTLIERS?  Extreme outliers are to left or right Q3  Mild outliers are between to left of Q1 or right of Q3  WHAT DO YOU DO WITH OUTLIERS?  Leave them in data if it is important  Remove them if they are different from rest  Correct error in data entry  **DATA MODELLING; DIST. FITTING**: Sample data limits our ability to predict uncertain events; potential values outside range of sample data are not included; better to identify underlying probability dist. from which sample data come by “fitting” theoretical dist. to data; verifying goodness of fit statistically  **PROB. DIST.:** characterization of possible values that random variable may assume along with probability of assuming these values  3 PERSPECTIVES FOR DEVELOPING  • theoretical arguments  • empirical data – empirical probability dist.  • using subjective values; expert judgement  Why do we need to know about distribution?  Helps you to understand underlying process that generates sample data. Useful in building decision models with theoretical dist. of data. Helps to compute prob. of occurrence of outcomes to assess risk; make decisions  **GOODNESS OF FIT:** fitting data to probability dist.  CHI-SQUARE (need at least 50 data points)  KOLMOGOROV-SMIRNOV (works well for small samples; only for non-parametric data)  ANDERSON-DARLING (puts more weight on differences between tails of dist.)  SHAPIRO’S WILKINS NORMALITY TEST (test data against normal dist.) 🡪 P-value > 0.05 implies that dist. of data is not significantly different from normal dist. In other words, we can assume data is normal.  **STANDARD NORMAL**: X is called as standard normalrandom variable when ; ; Z,            **NORMAL DIST.:**  is bell-shaped curve  Remark: **1**. Symmetric; **2**. Mean = Median = Mode;  **3**. Range of X is unbounded;  **4.** Empirical rules apply (i.e., area under density function within ± 2 standard deviation is 95.4%, within ±3 standard deviation is 99.7%) | **POISSON BINOMIAL:** Let . When ; remains a constant as .  The approximation is good when  OR . If is close to 1, we can still use Poisson distribution to approximate binomial probabilities.  **NORMAL BINOMIAL**  Use when:  When is small; is not extremely close to 0 or 1, approximation is fairly good.  Use normal approximation only if  **CONTINUITY CORRECTION:**  is binomial random variable mean , .  **GAMMA FUNCTION:** ( is a complex number with positive real part). Gamma function is defined by    For integer,  **DISTRIBUTION:** chi-square or distribution with n degree of freedom is distribution of a sum of square of independent standard random variables .  Let , then  For large n, approximately.  If are independent chi-square random variables with degree of freedom  has distribution with degrees of freedom.  **– TABLE:** table contains values of for various n:  **STUDENT DISTRIBUTION:**    .  The t-table *shows*  In table degree of freedom  If random sample was selected from a normal population  are independent, so are  **FISHERS F-DISTRIBUTION:** *(ratio between two estimate of var.).* Random **samples** of size are selected from 2 **normal** **population** with variances  Table F-distribution gives value of such that | **STATISTICAL SAMPLING**  SAMPLING: foundation of statistical analysis.  ESTIMATORS: measures used to estimate unknown population parameters  POINT ESTIMATE: single number derived from sample that is used to estimate value of population parameters  **UNBIASED ESTIMATOR:** Let be estimator of (random var. based on sample). If , is unbiased estimator of  is an unbiased estimator of  An unbiased estimator of is  **A SAMPLING EXPERIENCE:** Sample size increases, average of sample means is all still close to expected value;  Standard deviation of sample means becomes smaller, meaning that means of samples are clustered closer together around true expected value.  **SAMPLING DIST.:** Sampling dist. of mean is dist. of means of all possible samples of fixed size n from some population.  Standard deviation of sampling dist. of mean is called **STANDARD ERROR** of mean  🟑As n, standard error, sampling error .  **Sampling distribution related to sample MEAN:**  *Infinite population or from a finite population with replacement having mean ; variance , sample distribution of sample mean has mean; variance is:*  **Sampling distribution relate to sample VAR:**  Let S2 be sample variance of a random sample of size n taken from a normal population with  is degrees of freedom.  **SAMPLING ERROR:**  SAMPLING (STATISTICAL) ERROR: samples are only subset of total population  SAMPLING ERROR depends on size of sample relative to population.  NON-SAMPLING ERROR: sample does not adequately represent target population, results from poor sample design or choosing wrong population frame. (e.g., convenience sample)  **EMPIRICAL RULES:** For normally distributed data set, proportion of values that lie within k  standard deviations of mean follow empirical rules:  Application of Empirical Rule - Process Capability Index *is measure of how well manufacturing process can achieve specifications*  Using sample of output, measure dimension of interest; compute total variation using third empirical rule.  **Estimating sampling error empirical rule**  Empirical rule for 3 standard deviations away from mean, ~99.7% of sample mean should be between:    ;  🟑 n increases, standard error, sampling error . | **LAW OF LARGE NUMBER LLN:** Let be a random sample of size with mean ; variance . Then,  **CENTRAL LIMIT THEOREM:**  Let be a random sample of size with mean ; variance .  ***Normal distribution provides an excellent approximation to sampling distribution of mean if n ≥* 30*.***  If are (approximately) , then is (approximately) regardless of sample size .  If sample size is large enough, then sampling dist. of mean ~normally distributed regardless of dist. of population ~ sample mean = population mean  If population ~ normally distributed, sampling dist. ~ normal distr. for any sample size.  **Interval estimates:**  probability interval is any interval [A, B] such that probability of falling between; B is . probability intervals are centred on mean/ median.  **CONFIDENCE INTERVALS** is range of values between which value of population parameter is believed to be, along with probability that interval correctly estimates true (unknown) population parameter.  The interval computed is called confidence interval for . fraction is called confidence coefficient or degree of confidence  *For 95% confidence interval, if we chose 100 different samples, leading to 100 different interval estimates, we would expect that 95% of them would contain true population mean.*  🟑 Explain difference as level of confidence decreases from 95% to 90%.  When level of confidence decreases from 95% to 90%, range of CI 🡪 rejection area  **CI for with KNOWN :**  is number with an upper-tail probability of for standard normal distribution Z.  **SAMPLE SIZE FOR ESTIMATING :** For margin of error ,    **CONFIDENCE INTERVALS; SAMPLE SIZE**  Determine appropriate sample size needed to estimate population parameter within specified level of precision . |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **CONFIDENCE INTERVALS FOR SPECIAL CASES:**   |  |  |  | | --- | --- | --- | | **CI for PROPORTION*:*** Let (*sample proportion*), *where is number in sample having desired characteristic; is sample size.*  *Suppose that we wish to determine number of voters to poll to ensure sampling error of at most ± 2%. With no information, use = 0.5 (proportion who poll):*  Use sample proportion from preliminary sample as estimate of or set = 0.5 | | | | **CI of ; KNOWN with**  **NORMAL population or** | |  | | **CI of ; UNKNOWN**  **NORMAL population &** | |  | | **CI of ; UNKNOWN with**  **NORMAL population or** | |  | | **CI of ; KNOWN**  **NORMAL population or** | |  | | **CI of ; UNKNOWN**  **with** | |  | | **CI of ; UNKNOWN ; NORMAL population;**  ***:***Define  *is pooled sample variance, an estimator for*  *confidence interval for is* | | | | **CI of ; KNOWN**  **with** |  | | | **CI of with NORMAL population,** |  | | | **CI of with** |  | | | **CI of ; KNOWN with**  **NORMAL population** |  | | | **CI of ; UNKNOWN with NORMAL population** |  | | | **CI of ; UNKNOWN : NORMAL population** | | |   **PAIRWISE ASSOCIATION TEST:** On left is used non-parametric test of pairwise correlation (robust to non-normal population/samples)    **Reference for statistical analysis using SAS, Stata, SPSS, R**  [https://stats.idre.ucla.edu/other/mult-pkg/whatstat/#](https://stats.idre.ucla.edu/other/mult-pkg/whatstat/) | **USING C.I. FOR DECISION MAKING:**  **1**. Required volume for bottle-filling process is 800; sample mean is 796 mls. We obtained confidence interval for population mean of [790.12, 801.88]. Should machine adjustments be made? Although sample mean is less than 800, sample does not provide sufficient evidence to draw that conclusion that population mean is less than 800 because 800 is contained within confidence interval.  **2**. 1,300 voters found that 692 voted for candidate in two-person race. This represents proportion of 53.23% of sample. Could we conclude that candidate will likely win election? 95% confidence interval for proportion is [0.505, 0.559] This suggests that population proportion of voters who favour this candidate is highly likely to exceed 50%, so it is safe to predict winner.  **3.** What if sample proportion is 0.515; confidence interval for population proportion is [0.488, 0.543]? Even though sample proportion is larger than 50%, sampling error is large; confidence interval suggests that it is reasonably likely that true population proportion could be less than 50%, so you cannot predict winner.  **PREDICTION INTERVALS** is 1 that provides range for predicting value of new observation from same population.  While confidence interval is associated with sampling dist. of statistic, but prediction interval is associated with dist. of random variable itself.  **CONFIDENCE interval VS PREDICTION interval:**   * *For 95% confidence interval means if we randomly choose n samples, there is 95% chance of them having desired values in ranges of 95% confidence level* * *For 95% prediction interval means that when there is new data coming, there is 95% chance that having desired values in 95% prediction level*   **HYPOTHESIS TESTING**  Null hypothesis: What you do not want to see  Alternative hypothesis: what you want to see  HYPOTHESIS TESTING PROCEDURE:  Step 1: Null; Alternative hypothesis (what you want to test). equal part is always in null hypothesis  Step 2: Determine level of significance (); power ().   |  |  |  | | --- | --- | --- | |  |  |  | | Reject | Type I error | Correct decision  = | | Not reject | Correct decision  P(not rejecttrue) | Type II error  P(not rejectfalse) |   Step 3: *Identify test statistic, distribution; rejection criteria.*    Step 4: Compute test statistic value based on your data.  Step 5: Conclusion.  **VALUE:** do not reject , else reject  **IMPROVING POWER OF TEST**  Power of test  • probability of not committing type II error  • should be high to make valid conclusion  How to ensure sufficient power?  🟑Power of test is sensitive to sample size  🟑small sample sizes 🡪 low power  🟑large sample required for small ⍺ | **HYPOTHESIS *on* ; KNOWN :NORMAL population or :**  To test:  When is true, we have test statistic:    **HYPOTHESIS on ; UNKNOWN NORMAL population**  To test:  When is true, we have test statistic:    **TWO-SIDED TEST ~ CONFIDENCE INTERVAL:**  confidence interval contains , is not located within rejection region will NOT be rejected.  **HYPOTHESIS on with KNOWN NORMAL population or :**  To test:  When is true, we have test statistic:  **HYPOTHESIS TEST ON :** To test:  We can use test statistic   |  |  | | --- | --- | |  | Rejection region | |  |  | |  |  | |  | ; |   **HYPOTHESIS on with UNKNOWN**    To test:  When is true, we have test statistic:  **HYPOTHESIS on with UNKNOWN NORMAL population;**  To test:  When is true, we have test statistic:  **HYPOTHESIS TEST ON PAIR SAMPLES:**  For paired sample, define:  To test:  When is true, we have test statistic:  **HYPOTHESIS TEST ON :**  To test:  We can use test statistic   |  |  | | --- | --- | |  | Rejection region | |  |  | |  |  | |  | or |   **SAMPLE TEST ON PROPORTION:**  **Calculator with Normal distribution:** | **REJECTION REGION; P-VALUE for**  **NORMAL distribution**:   |  |  |  | | --- | --- | --- | |  | *Rejection region* | *p-value* | |  |  |  | |  |  |  | |  | or |  |   **distribution:**   |  |  |  | | --- | --- | --- | |  | *Rejection region* | *p-value* | |  |  |  | |  |  |  | |  | or |  |   **T-TEST:** Paired two-sample for means  **TEST for EQUALITY of VARIANCES**  between 2 samples using new type of test, F-test.  • To use this test, we must assume that both samples are drawn from normal populations.  •  • F-test statistic:  **F-DIST.** has two degrees of freedom, 1 associated with numerator of F-statistic, ; 1 associated with denominator .  *Population with larger variance will be assigned numerator*  **ANALYSIS OF VARIANCE (ANOVA):** Used to compare means of two or more population groups; robust to departures from normality  at least 1 mean is different from others  • ANOVA measures variation between groups relative to variation within groups.  • Test statistic has F-dist. so if F-statistic is large enough based on level of significance chosen; exceeds critical value, we would reject .  **ANOVA Assumptions:** Independence, Normality; homogeneity of variances:  **1.** Randomly; independently obtained (validated if random samples are chosen)  **2.** Normally distributed;  **3.** Have equal variances  If sample sizes are equal, violation of third assumption does not have serious effects, but with unequal sample sizes, it can.  Comparing sample means of two populations, use t-test rather than ANOVA  **PEARSON CORRELATION TEST:**  *Test association between 2 quantitative variables:*  The test calculates Pearson correlation coefficient; p-value for testing non-correlation. Let x; y be two quantitative variables, where n samples are observed. *linear regression coefficient is*  Under , test statistic follows Student distribution with degree of freedom  **NONPARAMETRIC TEST OF PAIRWISE ASSOCIATION:** When to use it? Observe data distribution: presence of outliers; distribution of residuals is not Gaussian.  **SPEARMAN RANK-ORDER CORRELATION (quantitative ~ quantitative):** measure of monotonicity of relationship between two datasets  Like other correlation coefficients, this one varies between with 0 implying no correlation.  Correlations of or imply an exact monotonic relationship.  Positive correlations imply that as  Negative correlations imply that as | **Wilcoxon signed-rank test (quantitative ~ )**  The Wilcoxon signed-rank test is a non-parametric statistical hypothesis test used when comparing two related samples, matched samples, or repeated measurements on a single sample to assess whether their population mean ranks differ (i.e. it is a paired difference test). It is equivalent to one-sample test of difference of paired samples.  It can be used as an alternative to paired Student’s t-test, t-test for matched pairs, or t-test for dependent samples when population cannot be assumed to be normally distributed.  It has lower sensitivity compared to t-test. May be problematic to use when sample size is small  Null hypothesis difference between pairs follows a symmetric distribution around zero.  **MANN- WHITNEY TEST (quantitative ~ categorial 2 level):** also called Mann-Whitney-Wilcoxon / Wilcoxon rank-sum test / Wilcoxon-Mann-Whitney test is a nonparametric test of null hypothesis that two samples come from same population against an alternative hypothesis, especially that a particular population tends to have larger values than other.  It can be applied on unknown distributions contrary to e.g. a t-test has to be applied only on normal distributions. |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Linear model Regression  **LINEAR MODEL:** Given n random samples () linear regression models relation between observations and independent variables is  The are regression coefficients  is intercept or bias  are residuals  **An INDEPENDENT variable** also called **predictors.** It is a variable that stands alone and isn’t changed by other variables you are trying to measure. For example, someone’s age might be an independent variable. Other factors (such as what they eat, how much they go to school, how much television they watch) aren’t going to change a person’s age. In fact, when you are looking for relationship between variables you are trying to see if independent variable causes change in other variables, or dependent variables.  **A DEPENDENT variable,** called a **target variable**. It is something that depends on other factors. For example, a test score could be a dependent variable because it could change depending on several factors such as how much you studied, how much sleep you got night before you took test, or even how hungry you were when you took it. Usually when you are looking for a relationship between two things you are trying to find out what makes dependent variable change way it does.  **STANDARD ERROR** variability between observed; predicted Y values. This is formally called standard error of estimate, .  Step **1**: model data:  slope or coefficients or parameter of model  intercept or bias is second parameter of model  error, or residual with  Step **2**: fit: estimate model parameters. goal is to estimate  Minimises mean squared error MSE/Sum squared error SSE/Ordinary Least Squares OLS  **REGRESSION ANALYSIS** is tool for building mathematical; statistical models that characterize relationships between dependent (ratio) variable; 1 or more independent, or explanatory variables (ratio or categorical), all of which are numerical.  *Simple linear regression involves single independent variable* >< *multiple linear regression*  **RESIDUALS** are observed errors associated with estimating value of dependent variable using regression line:  Help detect outliers that bias regressions analysis.  Errors associated with individual observation  **RESIDUAL analysis; REGRESSION assumption**  Residual = Actual Y value Predicted Y value  Standard residual = residual/standard deviation  Rule of thumb: Standard residuals outside of orare potential outliers. | **LOSS FUNCTION:**  Ordinary Least Square(OLS): minimize the sum of squares of residuals**.**  **INTERACTIONS:** occurs when effect of 1 variable is dependent on another variable. We can test for interactions by defining new variable as product of two variables, ; testing whether this variable is significant, leading to alternative model.  Difference between correlation; interaction:  Whether two variables are associated says nothing about whether they interact in their effect on third variable. interaction between two variables means effect of 1 of those variables on third variable is not constant— effect differs at different values of other.  **REGRESSION STATISTIC: Multiple** ,  where r is sample correlation coefficient. r varies from (r is negative if slope is negative).  **F-TEST:**  **Goodness of fit:** of a statistical model describes how well it fits a set of observations. Measures of goodness of for typically summarizes discrepancy between observed values and values under model in equation. We will consider explained variance also known as co-efficient of determination, denoted  **(R-SQUARED)** is measure of “fit” of line to data.  A value of 1.0 indicates perfect fit; all data points would lie online; larger value of better fit.  value , order of polynomial ;  The total sum of squares is sum of squares explained by regression, plus sum of squares of residuals unexplained by regression, also called SSE such that .  The mean of y:  The total sum of squares, also called total squared sum of deviations from mean :  The regression sum of squares, also called explained sum of squares:  is estimated value of given a value of experience  The sum of squares of residuals, also called residual sum of squares RSS is:  is explained sum of squares of errors. It is variance by regression divided by total variance  **ADJUSTED R-SQUARED** adjusts for sample size; number of X variables.  **Why use adjusted R Square?**  R-squared has additional problems that adjusted R-squared is designed to address.  Problem 1: Add predictor to model, R-squared increases, even if due to chance alone. It never decreases 🡪 model with more terms appear to have better fit simply because it has more terms.  Problem 2: If model has too many predictors; higher order polynomials, it begins to model random noise in data. This is known as overfitting model; it produces misleadingly high R-squared values; lessened ability to make predictions. adjusted R-squared increases only if new term improves model more than would be expected by chance. It decreases when predictor improves model by less than expected by chance. adjusted R-squared can be negative, but it’s usually not, always | **Test:** Let be an estimator of variance of . two in denominator stems from two estimated parameters: intercept and coefficient.      The single degree of freedom comes from difference between  The fisher statistics of ratio of two variances:  Using F-distribution, compute probability of observing a value greater than F under with  : survival function (1Cumulative distribution function) at of given F-distribution  **Notice p-value (Significance F):** When p-value is less than threshold (significance level), justifies rejection of null hypothesis. Null hypothesis is rejected when p < 0.05; not rejected when p > 0 .05. Rejecting indicates X explains variation in Y  **INTERPRET REGRESSION ANALYSIS RESULT?**  **(e.g., )**  (y) Intercept (also called constant) = Mean value of Y if value of all  Coefficient of X = impact of X on Y;  t-value = statistical test on each coefficient; constant (0 or not)  F-statistic = statistical test on all coefficients; constant (all 0 or not)  Pseudo (for nonlinear regression; also, can be “adjusted”). Model fit in context of nonlinear models is usually defined in two ways  These are mathematically different from in linear models; therefore “Pseudo”  1) Degrees of improvement from intercept-only model - McFadden’s Pseudo  2) Use same idea of linear ; variance of Y explained by X - Pseudo - Pseudo works ok for simple regression where linearity assumption is not severely violated  Measure of fit for linear regression model. proportion of variance in dependent variable explained by exploratory variable(s) = 1 – (Sum of squares of residuals;  SSR / SST (Sum of square total) = Pearson Correlation Coefficient R (Y, Predicted Y from model) 2 = “R” squared!  SSR Unexplained variance of Y  SST =  2 = Total variance of Y  Adjusted (for linear regression): tend to increase as number of predictors in model increases. Adjusted calculates “accurate ” by penalizing with number of predictors; sample size  n = sample size, p = no. of predictors  Example: A study to predict number of home runs scored in a softball league with 32 teams of 9 players, based on different material used to make bat (alloy, composite, aluminium, hybrid); player’s experience playing in league. *What is appropriate number of independent variables for regression model?*  materials + year experience variables  **HOW TO FIND MODEL BEST FIT WITH DATA?**  **Method 1**. USE CHANGE (FOR NESTED MODELS, LINEAR REGRESSION): Find most parsimonious model by using ; Rationale - Parsimonious model is preferred if it fits data (at least) equal to more complex model; Two models are considered as “nested” if one is constrained version of other | 🡪 (2) nested in (1) because they are same if  🡪 (2) is more “parsimonious” than (1); estimate less no. of coefficients (= parameters)  We prefer Model (2) over Model (1) if change between Models (2); (1) are not statistically significant (= simpler but equally well fit data)  How to compare nested models in r?  **1)** Fit each Model (1); Model (2) 🡪 (model1) / (model2)  **2)** Use F-test to test change is statistically different 🡪 (model 1, model 2)  **3)** If test cannot reject H0 (= No difference) choose (2); otherwise stay with (1)  **Method 2**. USE INFORMATION CRITERION (IC; FOR NON-NESTED MODELS): Find best-performing model by using information criterion; Rationale - Best-performing model is preferred, considering its complexity; fit to data; Commonly used information criterion measures for model selection  ⎯ Akaike Information Criterion (AIC) or its adjusted version (AICc)  ⎯ Bayesian Information Criterion (BIC)  🡪 AIC/BIC are transformed values of a function of residuals; smaller is better  How to SELECT A MODEL by using information criterion in R?  1. Fit candidate models 🡪 e.g., …  2. For each model, calculate AIC or BIC 🡪 or use BIC 3. Choose model having smallest value of AIC or BIC  IMPORTANT!!!  ⎯ ICs are statistical measures; assume one candidate model is (close to) “TRUE” model  ⎯ “True model” - a model that represents true, exact relationship between Y; X(s)  ⎯ In practice, you CANNOT check this assumption; usually ok for multiple regression/time series  ⎯ You need to choose model fit measures most suitable for your model(s); data!!!  **MULTIPLE REGRESSION:**  Multiple linear regression is most basic supervised learning algorithm.  Given a set of regression, we assume model that generates data involves only a linear combination of input variables.  Extending each sample with an intercept allows us to use a more general notation based on linear algebra and write it as a simple dot product:  Where is a vector of weights that defined parameters of model. From now we have P and intercept.  Minimise Mean Squared Error MSE loss:  be an metric of N samples of P inputs features with one column of one and let be be a vector of N targets. mean squared error MSE loss is  The that minimise MSE can be found by: | **4 assumption of simple / multiple linear regression:**  1. Linearity (of relationship between Y & X): Residual vs. fitted - *Find straight horizontal line*  2. Normality of Errors = Errors (e; residuals) are normally distributed:  Normal Q-Q plot - Look for linear relationship  3. Homoscedasticity = Constant / Equal variance of errors (e) for all values of X = Impact of X on Y is same for all X values:  Residual vs. fitted; Scale-location - Look for straight horizontal line  4. Independence of errors = There is no correlation between errors (e) calculated from regression model - Need additional plot/test  \* Residual time series plot  \* Durbin-Watson test  • For cross-sectional data, this is usually not major issue  • Panel/time-series data need to check  • Issues 2, 3; 4 are often interrelated  • Cross-sectional data – data is collected only once, from different individuals/entities  • Panel/time-series data – data is collected multiple times from each individual/entity  **MULTIPLE regression categorical independent variable / factor:** **analysis of covariance (ANCOVA)**  Analysis of covariance (ANCOVA) is a linear model that blends ANOVA and linear regression. ANCOVA evaluates whether population means of a dependent variable (DV) are equal across levels of a categorical independent variable (IV) often called a treatment, while statistically controlling for effects of other quantitative or continuous variables that are not of primary interest, known as covariates (CV).  **Regression as analysis of variance:** ANOVA conducts F-test to determine whether variation in Y is due to varying levels of X.  ANOVA test for significance of regression:  : population slope coefficient = 0  : population slope coefficient ≠ 0  **One way AN(C)OVA:**  ANOVA: *one categorical independent variable, one factor*  ANCOVA: ANOVA with some co-variates.  **Two way AN(C)OVA:**  with two categorical independent variables, two factors.  **MULTIPLE COMPARISONS:** Note that under null hypothesis distribution of p-values is uniform.  **Statistical measures:**  True Positive (TP) equivalent to a hit. test correctly concludes presence of an effect.  True Negative (TN). test correctly concludes absence of an effect.  False Positive (FP)equivalent to a false alarm, Type I error. test improperly concludes presence of an effect. at < 0.05 leads to FP. False Negative (FN)equivalent to a miss, Type II error. test improperly concludes absence of an effect.  **correction for multiple comparisons**  The correction is based on idea that if an experimenter is testing 𝑃 hypotheses, then one way of maintaining familywise error rate (FWER) is to test each individual hypothesis at a statistical significance level of 1/𝑃 times desired maximum overall level.  So, if desired significance level for whole family of tests is 𝛼 (usually 0.05), then correction would test each individual hypothesis at a significance level of 𝛼/𝑃. For example, if a trial is testing 𝑃 = 8 hypotheses with a desired 𝛼 = 0.05, then correction would test each individual hypothesis at 𝛼 = 0.05/8 = 0.00625.  **The False discovery rate (FDR) correction for multiple comparisons**  FDR-controlling procedures are designed to control expected proportion of rejected null hypotheses that were incorrect rejections (“false discoveries”). FDR-controlling procedures provide less stringent control of Type I errors compared to familywise error rate (FWER) controlling procedures (such as correction), which control probability of at least one Type I error. Thus, FDR-controlling procedures have greater power, at cost of increased rates of Type I errors. | **BRAIN VOLUMES STUDY:** study provides brain volumes of grey matter (gm), white matter and cerebrospinal fluid) of 808 anatomical MRI scans.  1. Fetch demographic data demo.csv and tissue volume data ().  2. Merge tables.  3. Compute Total Intra-cranial volume.  4. Compute tissue ratios:  5. Descriptive analysis per site in excel file.  6. Visualize site effect of gm ratio using violin plot:  .  7. Visualize age effect of gm ratio using scatter plot:  8. Linear model ():  **OVERFITTING** In statistics and machine learning, overfitting occurs when a statistical model describes random errors or noise instead of underlying relationships.  Fitting model too closely to sample data at risk of not fitting it well to population in which we are interested.  Overfitting generally occurs when a model is **excessively complex**, such as having **too many parameters relative to number of observations**. A model that has been overfit will generally have poor predictive performance, as it can exaggerate minor fluctuations in data.  A learning algorithm is trained using some set of training samples. If learning algorithm has capacity to overfit training samples performance on **training sample set** will improve while performance on unseen **test sample set** will decline.  –value will increase if we fit higher order polynomial functions to data 🡪 make it difficult to explain phenomena rationally.  In multiple regression, if we add too many terms to model, then model may not adequately predict other values from population.  Overfitting can be mitigated by using good logic, intuition, theory; parsimony  🟑Overfitting can be prevented by adding more data  The overfitting phenomenon has **three main explanations**: excessively complex models, multicollinearity, and high dimensionality.  **REASON1: Model complexity**  Complex learners with too many parameters relative to number of observations may overfit training dataset. *We should keep model as simple as possible.*  PRINCIPLE OF PARSIMONY: Good models are as simple as possible  **REASON 2: Multicollinearity**  Predictors are highly correlated, meaning that one can be linearly predicted from others.  In this situation coefficient estimates of multiple regression may change erratically in response to small changes in model or data.  Alternatively, we can say that Multicollinearity occurs when there are strong correlations among independent variables, they can predict each other better than dependent variables.  Multicollinearity does not reduce predictive power or reliability of model as a whole, at least not within sample data set; it only affects computations regarding individual predictors. That is, a multiple regression model with correlated predictors can indicate how well entire bundle of predictors predicts outcome variable, but it may not give valid results about any individual predictor, or about which predictors are redundant with respect to others.  In case of perfect multicollinearity, predictor matrix is singular and therefore cannot be inverted.  Under these circumstances, for a general linear model , ordinary least-squares estimator, does not exist. |

|  |  |  |  |
| --- | --- | --- | --- |
| **Dealing with :**  Regularization by shrinkage: Introduce a bias in solution by making non-singular.  🟑Feature selection: select a small number of features.  There are many potential benefits of variable and feature selection: facilitating data visualization and data understanding, reducing measurement and storage requirements, reducing training and utilization times, defying curse of dimensionality to improve prediction performance.  🟑Feature selection: select a small number of features using shrinkage.  🟑Extract few independent (uncorrelated) features using principal components analysis (PCA), partial least squares regression (PLS-R) or regression methods to cut number of predictors to a smaller set of uncorrelated components.  **REASON 3: High dimensionality**  High dimensions mean a large number of input features.  Linear predictor associate one parameter to each input feature, so a high-dimensional situation (𝑃, number of features, is large) with a relatively small number of samples 𝑁 (so-called large 𝑃 small 𝑁 situation) generally lead to an overfit of training data. Thus, it is generally a bad idea to add many input features into learner. This phenomenon is called curse of dimensionality.  One of most important criteria to use when choosing a learning algorithm is based on relative size of 𝑃 and 𝑁.  🟑Remember covariance matrix used in linear model is matrix of rank . This if equation system is over parameterised and admit an infinity of solutions that might be specific to learning dataset. See also ill-conditioned or singular matrices.  The sampling density of N samples in a P-dimensional space is proportioned to . Thus, a high-dimensional space becomes very sparse, leading to poor estimations of samples densities.  Another consequence of sparse sampling in high dimensions is that all sample points are close to an edge of sample. Consider N data points uniformly distributed in a P-dimensional unit ball centred at origin. Suppose we consider a nearest-neighbour estimate at origin. median distance from origin to closest data point is given by expression:  A more complicated expression exists for mean distance to closest point. For , , more than halfway to boundary. Hence most data points are closer to boundary of sample space than to any other data point. reason that this present a problem is that prediction is much more difficult near edges of training sample. One must extrapolate from neighbouring sample points rather than interpolate between them.  🟑 Structural risk minimization provides a theoretical background of this phenomenon.  🟑 Bias-variance trade-off. | **REGULARISED REGRESSION:**  **REASON:**  Linear regression minimizes a loss function  It chooses a coefficient for each feature variable  Large coefficients can lead to overfitting  🡪 **Penalizing large coefficients: Regularization**  **RIDGE REGRESSION ( Regularization):**  **LOSS FUNCTION:**  Alpha is the parameter we need to choose  Picking alpha here is similar to picking k in KNN  **Alpha controls model complexity:**  If we get back to the OLS 🡪 high probability lead to the overfitting  If high , we can get under fitting  **ALTERNATIVELY, we can understand as below:**  Overfitting generally leads to excessively complex weight vectors, accounting for noise or spurious correlations within predictors. To avoid this phenomenon learning should constrain solution in order to fit a global pattern. This constraint will reduce (bias) capacity of learning algorithm.  Adding such a penalty will force coefficients to be small, i.e. Shrink them towards zeros.  Therefore, loss function generally SSE, is combined with a penalty function leading to form:  The respective contribution of loss and penalty is controlled by regularization parameter .  Ridge regression impose a penalty on coefficients, i.e. it penalises with Euclidean norm of coefficients while minimizing SSE. objective function becomes:  The that minimise can be found by following derivation:  The solution adds a positive constant to diagonal of before inversion. This makes problem non-singular, even if is not of full rank and was main motivation behind ridge expression.  Increasing shrinks coefficients toward 0.  This approach penalises objection function by Euclidean norm of coefficients such that solutions with large coefficients become unattractive.  This ridge penalty shrinks coefficients toward zero.  **Ridge Fisher classification (L2 regularization)** When matrix 𝑆𝑊 is not full rank or 𝑃 ≫ 𝑁, Fisher most discriminant projection estimates of is not unique. This can be solved using a biased version of :  where 𝐼 is 𝑃 × 𝑃 identity matrix. This leads to regularized (ridge) estimator of Fisher’s linear discriminant analysis:  Increasing 𝜆 will:  • Shrinks coefficients toward zero.  • covariance will converge toward diagonal matrix, reducing contribution of pairwise covariance  **RIDGE logistic regression (L2-regularization)**  The objective function to be minimized is now combination of logistic loss log (𝑤) with a penalty of L2 norm of weights vector. In two-class case, using 0/1 coding we obtain: | **LASSO REGRESSION ( Regularization)**  **Lasso** stands for Least Absolute Shrinkage and Selection Operator  **LOSS FUNCTION:**  **ALTERNATIVELY, we can understand as below:**  Lasso regression penalise coefficients by norm. This constraint will reduce (bias) capacity of learning algorithm. To add such a penalty forces coefficients to be small, it shrinks them toward zero. Objective function to minimize becomes:  Alternatively, we can express objective function as  This penalty forces some coefficients to be exactly zero, providing a feature selection property.  Minimizing the objective function is the same as minimizing the sum of squares with constraint . Some of the s are shrunk to exactly zero, resulting in a regression model that is greater to interpret,  **A tuning parameter, control strength of Lasso penalty. is basically amount of shrinkage.**  When no parameter are eliminated. The estimate is equal to the one found in linear regression.  As increases, more and more coefficients are set to zero and eliminated. Theoretically, when then all the coefficients are eliminated.  As increases bias increases.  As decreases, variance increases.  **SPARSITY OF NORM:**  **Principle of parsimony**  The simplest of two competing theories is to be preferred. Definition of parsimony: Economy of explanation in conformity with Occam’s razor.  Among possible models with similar loss, choose simplest one:  • Choose model with smallest coefficient vector, i.e. smallest or norm of 𝛽, i.e. or penalty. See also bias-variance trade-off.  • Choose model that uses smallest number of predictors. In other words, choose model that has many predictors with zero weights. Two approaches are available to obtain this: Perform a feature selection as a pre-processing prior to applying learning algorithm, or embed feature selection procedure within learning process.  **Sparsity-induced penalty or embedded feature selection with ℓ1 penalty**  The penalty based on ℓ1 norm promotes sparsity (scattered, or not dense): it forces many coefficients to be exactly zero. This also makes coefficient vector scattered. Figure bellow illustrates OLS loss under a constraint acting on ℓ1 norm of coefficient vector. I.e., it illustrates following optimization problem:  subject to  **Optimization issues:**  • No more closed-form solution.  • Convex but not differentiable.  • Requires specific optimization algorithms, such as fast-iterative shrinkage-thresholding algorithm (FISTA)  **LASSO REGRESSION for FEATURE SELECTION:**  Can be used to select important features of a dataset  Shrinks coefficients of less important features to exactly 0. | **LASSO logistic regression (L1-regularization)**  The objective function to be minimized is now combination of logistic loss log (𝑤) with a penalty of L1 norm of weights vector. In two-class case, using 0/1 coding we obtain:    Ridge linear Support Vector Machine (L2-regularization) Support Vector Machine seek for separating hyperplane with maximum margin to enforce robustness against noise. Like logistic regression it is a **discriminative method** that only focuses of predictions.  Here we present non-separable case of Maximum Margin Classifiers with coding (ie.: ). In next figure legend aply to samples of “dot” class. **Linear SVM for classification (also called SVM-C or SVC) minimizes**:      With  Here we introduced slack variables: , with = 0 for points that are on or inside correct margin boundary and for other points. Thus:  🟑If then point lies outside margin but on correct side of decision boundary. In this case = 0. constraint is thus not active for this point. It does not contribute to prediction.  🟑If 1 > (𝑤 · ) ≥ 0 then point lies inside margin and on correct side of decision boundary. In this case 0 < ≤ 1. constraint is active for this point. It does contribute to prediction as a support vector.  🟑If 0 < then point is on wrong side of decision boundary (misclassification). In this case 0 < > 1. constraint is active for this point. It does contribute to prediction as a support vector. This loss is called hinge loss, defined as:  So linear SVM is closed to Ridge logistic regression, using hinge loss instead of logistic loss. Both will provide very similar predictions.  **LASSO LINEAR SUPPORT VECTOR MACHINE (L1 REGULARIZATION)**  Linear SVM for classification (also called SVM-C or SVC) with l1-regularization    With  **Elastic-net classification (L2-L1-regularization)**  The objective function to be minimized is now combination of logistic loss log (𝑤) or hinge loss with combination of L1 and L2 penalties. In two-class case, using 0/1 coding we obtain:  **ELASTIC-NET REGRESSION (-regularization)**  The Elastic-net estimator combines and penalties, and results in problem to  where acts as a global penalty and as an ratio.  RATIONALE • If there are groups of highly correlated variables, Lasso tends to arbitrarily select only one from each group. These models are difficult to interpret because covariates that are strongly associated with outcome are not included in predictive model. Conversely, elastic net encourages a grouping effect, where strongly correlated predictors tend to be in or out of model together.  • Studies on real world data and simulation studies show that elastic net often outperforms lasso, while enjoying a similar sparsity of representation. |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Multivariate statistics  Multivariate statistics include all statistical techniques for analysing samples made of two or more variables. data set is a collection of independent samples column vectors of length P    PRE-REQUITSIVE: LINEAR ALGEBRA  **MEAN VECTORS:**  The mean column-vector whose estimator is  **COVARIANCE MATRIX:**  The co-variance matrix a symmetric positive semi-definite matrix whose element in position is co-variance between and  The co-variance matrix generalises notion of covariance to multiple dimensions  The co-variance matrix describes shape of sample distribution around mean assuming an elliptical distribution    Whose estimator matrix is given by  If we assume that X is centred, X is replaced by then estimator is  is an estimator of co-variance between variables **PRECISION MATRIX**  In statistics, precision is reciprocal(inverse) of variance, and precision matrix is matrix inverse of covariance matrix. It is related to **partial correlations** that measure degree of association between two variables, while controlling effect of other variables  **MAHALANOBIS DISTANCE**  Reference:<https://www.machinelearningplus.com/statistics/mahalanobis-distance/>  distance is an effective multivariate distance metric that measures distance between a point and a distribution. It is an extremely useful metric having, excellent applications in multivariate anomaly detection, classification on highly imbalanced datasets and one-class classification.  The distance is a measure of distance between two points x and (or we can say it is distance of a point to a distribution) where dispersion (the covariance structure) of samples is considered.  This dispersion is considered through covariance matrix. This is formally expressed as    is vector of observation (row in dataset)  is te vector of mean values of independent variables (mean of each column)  inverse covariance matrix of independent variables.  is essentially distance of vector from mean. We then divide this by covariance matrix (or multiply by inverse of covariance matrix).  If you think about it, this is essentially a multivariate equivalent of regular standardization (). That is, | **Intuitions** • Distances along principal directions of dispersion are contracted since they correspond to likely dispersion of points.  • Distances to principal directions of dispersion are since they correspond to unlikely dispersion of points.  If covariance matrix is identity matrix, distance reduces to Euclidean distance. If covariance matrix is diagonal, then resulting distance measure is called a normalized Euclidean distance.  More generally, distance is a measure of distance between a point x and a distribution . It is multi-dimensional generalization of idea of measuring how many standard deviations away x is from mean: along each principal component axis, it measures number of standard deviations from x to mean of distribution  *distance is different from Euclidean distance:*  1. It transforms columns into uncorrelated variables.  2. Scale columns to make their variance equal 1  3. Calculate Euclidean distance  **EUCLIDEAN distance with multivariate model:**  Euclidean distance is commonly used straight line distance between two points.  *Euclidean distance will work fine if dimensions are equally weighted and are independent of each other.*   |  |  |  |  | | --- | --- | --- | --- | | Area (sq.ft) | Price ($1000) | Area (acre) | Price ($M) | | 2400 | 1560000 | 0.0550944 | 156 | | 1950 | 126750 | 0.04476 | 126.75 |   The two table above show area and price of same objects. Only units of variables change.  Issue 1: Since both tables represent same entities, distance between any rows, point A and point B should be same. But Euclidean distance gives a different value even though distances are technically same in physical space.  This can be overcome by scaling variables, by computing or make it vary within a range between 0 and 1  Issue 2: if dimensions are correlated to of another, which is distribution) can give little or misleading information about how close a point really is to cluster   |  |  | | --- | --- | | https://www.machinelearningplus.com/wp-content/uploads/2019/04/Mahalanobis_Distance_Usecase.jpg | When X and Y are uncorrelated, Euclidean distance from centroid can be useful to infer if a point is member of distribution. farther it is, less likely it is a member. | | Both point A and point 2 have same Euclidean distance from centroid, but only point 1 is a member of distribution. To detect point 2 as outlier, (point 2, centroid) should be much higher than (point 1, centroid), distance can be used instead. | |   typically, case in real-world datasets, Euclidean distance between a point and centre of points (**Usage of Mahalanobis distance:**  Usage 1: Multi-covariate outlier detecting using Mahalanobis distance.  Usage 2: Mahalanobis distance for classification problem  Usage 3: One-class classification  One Class classification is a type of algorithm where training dataset contains observations belonging to only one class.  With only that information known, objective is to figure out if a given observation in a new (or test) dataset belongs to that class.  *For example*, you have a large dataset containing millions of records that are NOT yet categorized as 1’s and 0’s. But you also have with you a small sample dataset containing only positive (1’s) records. By learning information in this sample dataset, you want to classify all records in large dataset as 1’s and 0’s.  Based on information from sample dataset, it is possible to tell if any given sample is a 1 or 0 by viewing only 1’s (and having no knowledge of 0’s at all). | **MULTI-VARIATE NORMAL DISTRIBUTION:**  *The distribution, or probability density function PDF of continuous random variable is a function that describes relative likelihood for this random variable taken on a given time*  The multivariate normal distribution or multivariate Gaussian distribution of a P-dimensional random vector is |  |  |  |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Time series and Forecasting  **TIME SERIE** – stream of historical data, daily  Have components such as:  **1.** random behaviour;  **2.** trend: is gradual upward or downward movement of time series;  **3.** seasonal effects: is 1 that repeats at fixed intervals of time, typically year, month, …;  **4.** cyclical effects: describe ups; downs over much longer time frame, i.e. several years  **STATIONARY TIME SERIES** have only random behaviour.  A time series is said to be stationary is its statistical properties such as mean, variance remain constant over time: constant mean, constant variance, an auto-covariance does not depend on time  There are 2 reasons making a time series **non-stationary**:  Trend – varying mean over time  Resampling, smoothing, widowing, rolling average  for each time point, take average of points on either side of it. Note that number of points is specified by a window size  Seasonality – variations at specific timeframes.  First-order differencing  The  of a time series is series of changes from one period to next. If  denotes value of time series Y at period t, then first difference of Y at period t is equal to  Seasonal correlation: correlation of first-order differences of these time series  For example: diet’ and ‘gym’ are negatively correlated! Remember that you have a seasonal and a trend component. From correlation coefficient, ‘diet’ and ‘gym’ are negatively correlated:  • trends components are negatively correlated.  • seasonal components would positively correlate and theirs. *actual correlation coefficient is capturing both of those.*  **AUTOCORRELATION:**  A time series is periodic if it repeats itself at equally spaced intervals, say, every 12 months. Autocorrelation Function (ACF): It is a measure of correlation between TS with a lagged version of itself. For instance, at lag 5, ACF would compare series at time instant with series at instant ( and being end points).  ACF peaks every 12 months: Time series is correlated with itself shifted by 12 months.  Autocorrelation is present successive observations are correlated with 1 another; for example, large observations tend to follow other large observations; small observations also tend to follow 1 another.  In such cases, other approaches, called **autoregressive models**, are more appropriate.  **TIME series forecasting using autoregressive**  **moving average ARMA model:**  ARMA models are often used to forecast a time series. These models combine autoregressive and moving average models. In moving average models, we assume that a variable is sum of mean of time series and a linear combination of noise components.  The autoregressive and moving average models can have different orders. In general, we can define an ARMA model with p autoregressive terms and q moving average terms as follows: | **Choosing p and q**  Plot partial autocorrelation functions for an estimate of p, and likewise using autocorrelation functions for an estimate of q.  Partial Autocorrelation Function (PACF) measures correlation between TS with a lagged version of itself but after eliminating variations already explained by intervening comparisons. For example, at lag 5, it will check correlation but remove effects already explained by lags 1 to 4.  First, we draw autocorrelation and partial autocorrelation plots. In plot, you will see two dotted line on either side of 0 are confidence interval. These can be used to determine p and q values as:  : lag value where PACF chart crosses upper confidence intervals for first time.  : lag value where ACF chart crosses upper confidence interval for first time.  **FORECASTING TECHNIQUES:**  **Qualitative; Judgmental techniques** rely on experience; intuition.  **Historical analogy approach** obtains forecast through comparative analysis with prior situations;  **Delphi method** questions anonymous panel of experts 2-3 times in order to reach convergence of opinion on forecasted variable;  **Indicators** are measures that are believed to influence behaviour of variable we wish to forecast. Indicators are often combined quantitatively into **index**, single measure that weights multiple indicators, thus providing measure of overall expectation;  **Leading indicators:** series of measure change before variable change;  **Lagging indicators:** series of measures that follow change of variable.  **STATICALLY FORECASTING MODELS:**  **MOVING AVERAGE MODEL:** smoothing method based on idea of averaging random fluctuations in time series to identify underlying direction in which time series is changing. Simple moving average forecast for next period is computed as average of most recent k observations. Larger values of k result in smoother forecast models since extreme values have less impact  **EXPONENTIAL SMOOTHING MODEL:**  **Simple**  is called smoothing factor/ coefficient/ constant.  Value of dictates how much weight is given to most recent observed value versus last expected value;  regulates importance of most recent observations with respect to smoothed mean of previous values;  : *assign an almost constant weight to all past observations;*  : *assign an almost constant weight to all recent observations.*  **Double** *Rewrite simple exponential smoothing:*          Predicted value is a function of last estimates of level linear trend  modulates importance of most recent value of trend; with respect to trend of previous periods.  : *assign almost weight to trends in past*  : *most recently exhibited trend is pre-dominant.*  **REGRESSION-BASED FORECASTING FOR TIME SERIES WITH LINEAR TREND:**  Simple linear regression can be applied to forecasting using time as independent variable.  **Forecasting time series with SEASONALITY:** When time series exhibit seasonality, different techniques provide better forecasts than ones we have described:  Multiple regression models with categorical variables for seasonal components; | |  |  |  |  | | --- | --- | --- | --- | |  | **DISCRETE RANDOM VARIABLE** | | **CONTINUOUS RANDOM VARIABLE** | | ***Probability***  ***Mass/***  ***Density***  ***Function*** |  | |  | | ***Cumulative Distribution Function*** | If only *possible values are integers* and if *a and b are integers*, | |  | | ***Mean/***  ***Expectation/***  ***Expected values*** |  | |  | | *a; b constant, .* | |  | | ***Expectation/Mean***  ***of Function*** |  | |  | | is called moment of X is called second moment | | | | ***Variance*** |  | |  | | ***Joint Prob. Mass/***  ***Density Function*** |  | |  | | ***Marginal Distribution*** |  | |  | | ***Conditional Probability***  ***Mass***  ***Function*** |  | |  | | Conditional probability mass/density function of X:  X and Y **independent** for some x and y | |  | | *HOLT-WINTER MODEL*, similar to exponential smoothing models in that smoothing constants are used to smooth out variations in level; seasonal patterns over time.  **Holt-winter model for forecasting time series SEASONALITY; TREND:**  HOLT-WINTERS ADDITIVE MODEL applies to time series with relatively stable seasonality:  is smoothed estimate of level at time  is smoothed estimate of change in trend value at time  is smoothed estimate of appropriate seasonal component at  HOLT-WINTERS MULTIPLICATIVE MODEL applies to time series whose amplitude increases or decreases over time; is  **Regression forecasting with Causal variable:** In many forecasting applications, other independent variables besides time, i.e. economic indexes or demographic factors, may influence time series. Explanatory/causal models, often called econometric models, seek to identify factors that explain statistically patterns observed in variable being forecast, usually with regression analysis  **Practice of forecasting:** Judgmental; qualitative methods are used for forecasting sales of product lines; broad company; industry forecasts. Simple time-series models are used for short; medium-range forecasts. Regression methods are typically used for long term forecasts. | |  | | | **FORECASTING METHODS CLASSIFICATION:**   |  |  |  | | --- | --- | --- | |  | **NO SEASONALITY** | **SEASONALITY** | | ***No trend*** | Simple moving average or simple exponential smoothing | Holt-winter no-trend smoothing model or multiple regression | | ***Trend*** | Double exponential smoothing | Holt-winter addictive or Holt-winter multiplicative model. |  |  |  | | --- | --- | | **ERRORS METRICS; FORECAST ACCURACY:** *For all metrics, smaller values 🡪 better data* | | | Mean absolute deviation: focus on mean value of errors | Mean square error / deviation: focus on variance of errors | | Root mean square error focus on standard deviation of errors | Mean absolute percentage error: cannot be used if time series contains 0 (division by 0) |  |  |  |  | | --- | --- | --- | |  | **POPULATION OF SIZE N** | **SAMPLE OF N OBSERVATIONS** | | ***Mean*** |  |  | | ***Variance*** |  |  | | ***Co-var.*** |  |  | | ***Co-***  ***relation*** |  |  |  |  |  |  | | --- | --- | --- | |  | **PROBABILITY MASS/DENSITY FUNCTION** | **MEAN; VARIANCE** | | ***Discrete***  ***uniform***  ***distribution*** |  |  | | ***Continuous***  ***uniform***  ***distribution*** |  |  | | ***Bernoulli***  ***trials*** | *Experiment with 2 outcomes (“success”; “failure”)* |  | | ***Binomial***  ***distribution*** |  |  | | ***Negative***  ***binomial***  ***distribution*** | *#trials before obtain k successes;* |  | | ***Geometric***  ***distribution*** | Memoryless property of Geometric: | #*required trials until first success is achieved* | | #*failures until first success is achieved* |  | | ***Poisson***  ***random***  ***variable*** | *# success in fixed interval/period/region* |  | | ***Exponential***  ***distribution*** | Memoryless property of Exponential distribution: |  | | ***Normal***  ***distribution*** |  |  | | ***distribution*** | *degree of freedom* *gamma function*; |  | | ***Student’s***  ***distribution*** | *Z* *~*; U~ |  | | ***Fisher’s F***  ***distribution*** |  | | |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Dimension reduction feature extraction  In machine learning and statistics, dimensionality reduction or dimension reduction is process of reducing number of features under consideration and can be divided into feature selection (not addressed here) and feature extraction.  Feature extraction starts from an initial set of measured data and builds derived values (features) intended to be informative and non-redundant, facilitating subsequent learning and generalization steps, and in some cases leading to better human interpretations. Feature extraction is related to dimensionality reduction.  Input matrix X, dimension :  Where rows represent samples and columns represent variables.  The goal is to learn a transformation that extracts a few relevant features. This is generally done by exploiting co-variance input variables.  **SINGULAR VALUE DECOMPOSITION SVD and MATRIX FACTORIZATION:**  **MATRIX FACTORIZATION PRINCIPLES:**  Decompose data matrix into a product of a mixing matrix and a dictionary matrix  If we consider only a subset of components is approximated by a matrix  Each line of is a linear combination mixing of dictionary items V  dimensional data points lie in a space whose dimension is less than (2 dots lie on a line, 3 dots on a plane, …)  **SINGULAR DECOMPOSITION PRINCIPLES:**  Singular-value decomposition SVD factorises data matrix into a product; where  : right singular  is a orthogonal matrix  It is a **dictionary** of patterns to be combine (according to mixing coefficients) to reconstruct original samples.  **V** performs initial rotations (projection) along principal component directions, also called loadings  Each performs linear combination of variables that has maximum samples variance, subject to being un-correlated with previous  singular values  **D** is a diagonal matrix made of a singular value of X with  **D** scale projection along co-ordinate axes by  Singular values are square roots of eigenvalues of  : left-singular vectors  is an orthogonal matrix  Each row provides mixing coefficients of dictionary items to reconstruct samples  It may be understood that co-ordinates on new orthogonal basis (obtained after initial rotation) called principal components in PCA  **SVD FOR VARIABLES TRANSFORMATION:**  V transforms correlated variables (X) into a set of uncorrelated ones (UD) that better expose various relationships among original data items.  At same time, DVS is a method for identifying dimensions along with data points exhibit most variation | **PRINCIPAL COMPONENT ANALYSIS (PCA):**  **Principal components analysis** is man method used for linear dimension reduction  The idea of principal component analysis is to find K principal components direction (called **loadings**) that captures variation in data as much as possible  It converts set of dimernsional observations of possibly correlated variables into a set of dimensional samples . new variables are linearly uncorrelated. columns of are called **principal components**.  The **dimensional reduction** is obtained by using only components that exploit correlation (covariance) among original variables.  **PCA** is mathematically defined as an orthogonal linear transformation that transform data to a new co-ordinated system such that greatest variance by some projection of data comes to lie on first coordinate (called first principal component), second greatest variance on second coordinate and so on.  PCA can be thought of as fitting a P-dimensional ellipsoid to data, where each axis of ellipsoid represents a principal component. If some axis of ellipse is small, then variance along that axis is also small, and by omitting that axis and its corresponding principal component from our representation of that dataset, we lose only a commensurably small amount of information.  Finding K largest axes of ellipsoid will permit to project data onto a space having dimensionality while maximizing variance of projected data.  **DATA PREPROCESSING:**  **and scaling:** *These are both forms of pre-processing*numerical data*, that is, data consisting of numbers, as opposed to categories or strings*  🟑 a variable is subtracting mean of variable from each data point so that new variable's mean is 0.  Consider a data matrix X with column-wise zero empirical mean (the sample mean of each column has been shifted to zero) 🡪 X is replaced by  🟑 a variable is multiplying each data point by a constant in order to alter range of data.  All  means is scaling a dataset so that its minimum is 0 and its maximum 1. To achieve this, we transform each data point x to   is slightly different; it's job is to centre data around 0 and to scale with respect to standard deviation:  where μ and σ are mean and standard deviation of dataset, respectively. First note that these transformations merely change range of data and not distribution. You may later wish to use any other number of transforms, such as a log transform or a Box-Cox transform, to make your data look more Gaussian (like a bell-curve).  **Two main reasons for scaling** your data are  🟑Your predictor variables may have significantly different ranges and, in certain situations, such as when implementing k-NN, this needs to be mitigated so that certain features do not dominate algorithm;  🟑You want your features to be unit-independent, that is, not reliant on scale of measurement involved: for example, you could have a measured feature expressed in meters and I could have same feature expressed in centimetres. If we both scale our respective data, this feature will be same for each of us. | **EIGEN DECOMPOSITION of data COVAR. matrix:**  Consider projection onto a one-dimensional space . We can define direction off this space using a P-dimensional vector v, which for convenience (and without loss of generality) we shall choose to be a unit vector so that (note that we are only interested in direction defined by v, not in magnitude of v itself).  **PCA consists of two main steps:**  Projection in direction to capture greatest variance:  Each P-dimensional data point is then projected onto v, where co-ordinate (in co-ordinate system of v) is a scalar value, namely . We want to find vector v that maximise these coordinates along v, which we will see corresponds to maximise variance of projected data. This is equivalently expressed as  We can write this in matrix as  Where is a biased estimate of co-variance matrix of data with  We now maximise projected variance with respect to v. *Clearly, this has to be constrained maximization to prevent* .  To enforce this constraint, we introduce a multiplier, denoted by 🡪 make an unconstrained maximization of  By setting gradient with respect to v equal to zero, this quantity has a stationary point when  We note that is eigenvector of  If we left-multiply above equation by and make use of . Variance is given by  And so, variance will be at a maximum when v is equal to eigenvector corresponding to largest eigenvalue . This eigenvector is known as first principal component.  We can define additional principal in an incremental fashion by choosing each new direction to be what which maximises projected variance amongst all possible directions that are orthogonal to those already considered. If we consider general case of a K-dimensional projection space, optimal linear projection for which variance of projected data is maximised is not defined by K eigenvectors, , of data covariance matrix that correspond to K largest eigenvalues  **Back to SVD:** Sample covariance matrix of centred data X is:  Rewrite using SVD decomposition of X as  Considering only right singular vectors associated to singular value  Moreover, computing PCA with SVD do not require to form matrix , so computing SVD is not standard way to calculate principal components analysis from data matrix, unless only a handful of components are required.  **PCA outputs:** SVD or eigen-decomposition of data covariance matrix provides three main quantities:  🟑 Principal component directions (loadings) are eigenvectors of . or right-singular vectors of an SVD of X are called principal component directions of X. | 🟑 Principal components are in matrix C which is obtained by projecting X onto principal components’ directions.  Since and V is orthogonal  *Thus, . Hence, is simply projection of row vectors of X, input predictor vectors on direction , scaled by*  🟑The variance of each component is given by eigen values . It can be obtained from singular values:  **DETERMINING NUMBER OF PCS:**  We must choose number of required components this can be done by calculating explained variance ratio of first components and by choosing such that cumulative explained variance ratio is greater than some given threshold (). This is expressed as  cumulative explained variance  **INTERPRETATION and VISUALIZATION:**  **PCs:** Plot samples projected on first principal components as e.g. PC1 against PC2.  **PC directions**  Exploring loadings associated with a component provides contribution of each original variable in component.  Remark: loadings (PC directions) are coefficients of multiple regression of PC on original variables  Another way to evaluate contribution of original variables in each PC can be obtained by computing correlation between PCs and original variables, columns X, denoted , for . For PC, compute and plot correlations with all original variables.  These quantities are called **correlation loadings**.  **MULTI-DINMENSIONAL SCALING (MDS):**  The purpose of MDS is to find a low-dimensional projection of data in which pairwise distance between data points is preserved as closely as possible (in a least-square sense)  Let D be pairwise distance matrix where is a distance between point and  The MDS concept can be extended to a wide variety of data types specified in terms of a similarly matrix.  Given dissimilarity (distance) matrix , MDS attempts to find K-dimensional projections of N points, , concatenated in an matrix, so that are as close as possible. This can be obtained by minimization of a loss function called stress function  This loss function is known as Least-Square or Kruskal-Shepard scaling.  A modification of least-square scaling is  *mapping*.  The  *mapping* performs better at preserving small distances compared to Least-square scaling. | **CLASSICAL MULTI-DIMENSIONAL SCALING:**  Also known as principal coordinates analysis,  The distance matrix, D, is transformed to a similarly matrix, B, often using centred inner products.  The loss function becomes  The stress function in classical MDS is sometimes called *strain*.  • solution for classical MDS problems can be found from eigenvectors of similarity matrix.  • If distances in **D** are Euclidean and double centred inner products are used, results are equivalent to PCA  Example: dataset provides road distances (in kilometres) between 21 cities in Europe. Given this matrix of pairwise (non-Euclidean) distances , MDS can be used to recover coordinates of cities in some Euclidian referential whose orientation is arbitrary.  **DETERMINING NUMBER OF COMPONENTS:**  We must choose number of required components. Plotting values of stress function, obtained using components. In general, start with . Choose where you can clearly distinguish an elbow in stress curve.  This in this plot below, we choose to retrain information accounted for by first two components, since this is where elbow is in stress curve.  **NON-LINEAR DIMENSIONALITY REDUCTION:**  Non-linear dimensionality reduction or manifold learning cover unsupervised methods that attempt to identify low-dimensional manifolds within original P-dimensional space that represent high data density. Then those methods provide a mapping from high-dimensional space to low-dimensional embedding.  **ISOMAP** Isomap is a nonlinear dimensionality reduction method that combines a procedure to compute the distance matrix with MDS. distances calculation is based on geodesic distances evaluated on neighbourhood graph:  1. Determine neighbours of each point. All points in some fixed radius or K nearest neighbours.  2. Construct a neighbourhood graph. Each point is connected to other if it is a K nearest neighbour. Edge length equal to Euclidean distance.  3. Compute shortest path between pairwise of points to build distance matrix D.  4. Apply MDS on D. |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Data mining Clustering  **PARALLEL CO-ORDINATES CHART** consists of set of vertical axes, 1 for each variable selected. For each observation, line is drawn connecting vertical axes. point at which line crosses axis represents value for that variable.  **SCATTERPLOT matrix** combines several scatter charts into 1 panel, allowing user to visualize pairwise relationships between variables.  A **VARIABLE PLOT** plots matrix of histograms for variables selected.  **DIRTY DATA:** Real data sets that have missing values or errors, are called “dirty”; need to be “cleaned” before analysing them.  🟑Approaches for handling missing data.  🟑Eliminate records that contain missing data  🟑Estimate reasonable values for missing observations, i.e. mean or median value  🟑Use data mining procedure to deal with them. has capability to deal with missing data in Transform menu in Data Analysis group.  🟑Try to understand whether missing data are simply random events or there is logical reason.  🟑Eliminating sample data indiscriminately could result in misleading information; conclusions about data  **CLUSTER ANALYSIS**, or **DATA SEGMENTATION,** is collection of techniques that seek to group or segment collection of objects (observations or records) into subsets or clusters, such that those within each cluster are more closely related to 1 another than objects assigned to different clusters.  🟑Objects within clusters should exhibit high amount of similarity, whereas those in different clusters will be dissimilar.  Clustering is one of main tasks of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics.  **The goals of clustering** is to determine intrinsic grouping in a set of unlabelled data.  **A good clustering:** It can be shown that there is no absolute “best” criterion which would be independent of final aim of clustering. Consequently, it is user which must supply this criterion, in such a way that result of clustering will suit their needs.  **The algorithm is composed of following steps:**  1. Place K points into space represented by objects that are being clustered. These points represent initial group centroids.  2. Assign each object to group that has closest centroid  3. When all objects have been assigned, recalculate positions of K centroids  4. Repeat steps 2 and 3 until centroids no longer move | **K-MEANS CLUSTERING:**  Suppose we have a dataset that consists of N observations of random D-dimensional Euclidian variable . Our goal is to partition dataset into some number, K, of clusters, where we shall suppose for moment that value of K is given. Intuitively, we might think of a cluster as comprising a group of data points shoes inter-point distances are small compared to distances to points outside of cluster. We can formalise this notion by first introducing a set of D-dimensional vectors in which is a prototype associated with cluster. As we shall see shortly, we can think of as representing centres of clusters. Our goal is then to find an assignment of data points to clusters, as well as s set of vectors {} such that sum of squares of distances of each data points to its closest prototype vector is at a minimum.  It is convenient at this point to define some notation to describe assignment of data points to clusters. For each data point we introduce a corresponding set of binary indicator variables where that describe which of K clusters data point is assigned to, so that if data point is assigned to, so that if data point is assigned to cluster k then and for .  This is known as 1-of-K coding scheme, wen then define an objective function, denoted **inertia**, as:  Which represents sum of squares of Euclidian distances of each data point to its assigned vector . Our goal is to find values for to minimise function J. We can do this through an iterative procedure in which each iteration involves two successive steps corresponding to successive optimizations with respect to . | First, we choose some initial values for . Then in first phase we minimise J with respect to , keeping fixed. In second phase we minimise J with respect to , keep fixed. This two-stage optimization process is then repeated until convergence. We shall see these two stages of expectation-maximization EM algorithm, and to emphasize this we shall use terms E step and M step in context of K-means algorithm.  Consider first determination of . Because K in is a linear function of , this optimization can be performed easily to give a closed form solution. terms involving different are independent and so we can optimise for each separately by choosing to be 1 for whichever value of gives minimum value of . In other words, we simply assign data point to closest cluster centre. More formally, this can be expressed as  Now consider optimization of with held fixed. objective function J is a quadratic function of and it can be minimised by setting its derivative respect to to zero giving.  Which we can easily solve for  The denominator in this expression is equal to number of points assigned to cluster k, and so this result has a simple interpretation, namely set equal to mean of all data points assigned to cluster k. for this reason, procedure is known as K-means algorithm.  The two phases of re-assigning is equal to number of points assigned to clusters and re-computing cluster means are repeated in turn until there is no further change in assignments (or until some maximum number of iterations is exceeded). Because each phase reduces value of objective function J convergence of algorithm is assured. However, it may converge to a local rather than global minimum of J.  **HIERARCHICAL CLUSTERING**, data are *not* *partitioned* into *cluster* in single step. Instead, series of partitions takes place, which may run from single cluster containing all objects to n clusters, each containing single object.  Hierarchical clustering may be represented by two-dimensional diagram known as , which illustrates fusions or divisions made at each successive stage of analysis.  **AGGLOMERATIVE clustering methods** proceed by series of fusions of n objects into groups.  A BOTTOM-UP strategy, where each observation starts in their own cluster, and pairs of clusters are merged upwards in hierarchy.  **DIVISIVE clustering methods** separate n objects successively into finer groupings.  A TOP DOWN strategy, where all observations start out in same cluster, and then clusters are split recursively downwards in hierarchy.  Conclusion: In order to decide which clusters to merge or to split, a measure of dissimilarity between clusters is introduced. More specific, this comprise a distance measure and a linkage criterion. distance measure is just what it sounds like, and linkage criterion is essentially a function of distances between points, for instance minimum distance between points in two clusters, maximum distance between points in two clusters, average distance between points in two clusters, etc. One linkage criterion, Ward criterion, will be discussed next*.* | **AGGLOMERATIVE CLUSTERING METHODS:**  SINGLE LINKAGE CLUSTERING (nearest neighbour): Distance between groups is defined as distance between closest pair of objects, where only pairs consisting of 1 object from each group are focused. At each stage, closest 2 clusters are merged  COMPLETE LINKAGE CLUSTERING: distance between groups is distance between most distant pair of objects, 1 from each group  AVERAGE LINKAGE CLUSTERING: Uses mean values for each variable to compute distance between clusters  WARD’S HIERARCHICAL CLUSTERING: Uses sum of squares criterion  Ward clustering belongs to family of agglomerative hierarchical clustering algorithms. This means that they are based on a “bottoms up” approach: each sample starts in its own cluster, and pairs of clusters are merged as one moves up hierarchy.  In Ward clustering, criterion for choosing pair of clusters to merge at each step is minimum variance criterion. Ward’s minimum variance criterion minimizes total within cluster variance by each merge.  To implement this method, at each step: find pair of clusters that leads to minimum increase in total within-cluster variance after merging. This increase is a weighted squared distance between cluster centres.  The main advantage of agglomerative hierarchical clustering over 𝐾-means clustering is that you can benefit from known neighbourhood information, for example, neighbouring pixels in an image.  **GAUSSIAN MIXTURE MODELS:**  The Gaussian mixture model GMM is a simple linear superposition of Gaussian components over data, aimed at providing a rich class of density models. We turn to a formulation of Gaussian mixtures in terms of discrete latent variables: K hidden classes to be discovered.  **GAUSSIAN MIXTURE MODEL VS K-MEANS:**  🟑 Whereas 𝐾-means algorithm performs a hard assignment of data points to clusters, in which each data point is associated uniquely with one cluster, GMM algorithm makes a soft assignment based on posterior probabilities.  🟑Whereas classic 𝐾-means is only based on Euclidean distances, classic GMM use a Mahalanobis distances that can deal with non-spherical distributions. It should be noted that Mahalanobis could be plugged within an improved version of 𝐾-Means clustering. Mahalanobis distance is unitless and scale invariant and considers correlations of data set.  The Gaussian mixture distribution can be written as a linear superposition of K Gaussians in form:  Where are mixing coefficients also known as class probability of class k, they sum to one:  is conditional distribution of given a class . It is multi-variate Gaussian distribution defined over a P-dimensional vector of continuous variables.  The goal is to maximise log-likelihood of GMM: | To compute classes parameters: we sum over all samples, by weighting each sample by its responsibility or contribution to class : such that for each point its contribution to all classes sum to one . This contribution is conditional probability of class given (sometimes called posterior). It can be computed using Bayes’ rule:  Since class parameters, depend on responsibilities and responsibilities depend on class parameters, we need a two-step iterative algorithm: expectation-maximization EM algorithm.  **The expectation-maximization (EM) algorithm for Gaussian mixtures**  Given a Gaussian mixture model, goal is to maximize likelihood function with respect to parameters (comprised of means and covariances of components and mixing coefficients)  Initialise means mixing coefficients and  **E-step:** For each sample , evaluate parameters using current responsibilities for each class k using current parameters values.    **M-step:** For each class, re-estimate parameters using current responsibilities:  Evaluate log-likelihood:  And check for convergence of either parameters or log-likelihood. If convergence criterion is not satisfied return to step 1.  **MODEL SELECTION:**  **Bayesian information criterion:**  In statistics, Bayesian information criterion (BIC) is a criterion for model selection among a finite set of models; model with lowest BIC is preferred. It is based, in part, on likelihood function and it is closely related to Akaike information criterion (AIC). |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Classification  **CLASSIFICATION METHODS** seek to classify categorical outcome into 1 of two or more categories based on various data attributes.  🟑For each record in database, we have categorical variable of interest; number of additional predictor variables.  🟑For given set of predictor variables, we would like to assign best value of categorical variable.  **MEASURING CLASSIFICATION:** Find probability of making misclassification error; summarize results in classification matrix, which shows number of cases that were classified either correctly or incorrectly.  **USING TRAINING; VALIDATION DATA:**  Data mining projects typically involve large volumes of data. data can be partitioned into:  ▪ training data set – has known outcomes; is used to “teach” data-mining algorithm  ▪ validation data set – used to fine-tune model  ▪ test data set – tests accuracy of model  **CLASSIFYING NEW DATA:** after classification scheme is chose; best model is developed based on existing data, we use predictor variables as inputs to model to predict output  **LINEAR CLASSIFICATION:**  Given training set of 𝑁 sample, where is a multidimensional input vector with dimension 𝑃 and class label (target or response) (binary classification problem).  The vector of parameters 𝑤 performs a linear combination of input variables, 𝑥𝑇 𝑤. This step performs a projection or a rotation of input sample into a good discriminative one-dimensional subspace.  This score ( decision function) is transformed, using nonlinear activation function 𝑦(.), to a “posterior probabilities” of class 1 ,  𝑤ℎ𝑒𝑟𝑒 .  The decision surfaces correspond to (𝑥) = constant, so that constant and hence decision surfaces are linear functions of 𝑥, even if function 𝑓(.) is nonlinear.  A thresholding of activation provides predicted class label.  The vector of parameters, that defines discriminative axis, minimizes an objective function 𝑓 that is a sum of loss function and some penalties on weights vector Ω(𝑤)  **FISHER’S LINEAR DISCRIMINANT WITH EQUAL CLASS COVARIANCE:**  This geometric method does not make any probabilistic assumptions, instead it relies on distances. It looks for **linear projection** of data points onto a vector, , that maximizes between/within variance ratio, denoted . Under a few assumptions, it will provide same results as linear discriminant analysis (LDA), explained below. Suppose two classes of observations, and , have means and and same total within-class scatter (“covariance”) matrix    Where is matrix of data centred on their respective means: , where are matrices of samples of classes . Let being scatter “between-class" matrix given by  The linear combination of features have means for and variance . Fisher defined separation between these two distributions to be ratio of variance between classes to variance within classes: | **The Fisher most discriminant projection**  In two-class case, maximum separation occurs by a projection on using Mahalanobis metric , so that  Demonstration: Differentiating with respect to w gives  Since we do not care about magnitude of , only its direction, we replaced scalar factor by .  In multiple-class case, solutions are determined by eigenvectors of that correspond to largest eigenvalues. However, in two-class case ( it is easy to show that is unique eigenvector of  where here . Which leads to result *.*  **The separating hyperplane:**  The separating hyperplane is a *𝑃 -* 1-dimensional hyper surface, orthogonal to projection vector, *𝑤*. There is no single best way to find origin of plane along *𝑤*, or equivalently classification threshold that determines whether a point should be classified as belonging to or to . However, if projected points have roughly same distribution, then threshold can be chosen as hyperplane exactly between projections of two means, i.e. as  **DISCRIMINANT ANALYSIS** is technique for classifying set of observations into predefined classes*.*  Uses predefined classes based on set of linear discriminant functions of predictor variables  Based on training data set, technique constructs set of linear functions of predictors, known as discriminant functions:  are discriminant coefficients (weights), are input variables (predictors), c is constant (intercept)  MAXIMUM NUMBER OF FUNCTIONS = number of groups, or number of variables in analysis, whichever is smaller.  🟑The weights of determining discriminant functions are computed by maximizing variance between groups relative to variance within groups.  For new observation, each of discriminant functions is evaluated; observation is assigned to class if discriminant function has highest value.  **LINEAR DISCRIMINANT ANALYSIS:** is a probabilistic generalization of Fisher’s linear discriminant. It uses Bayes’ rule to fix threshold based on prob. of classes.  Step 1: First, compute class-conditional distributions of given class , where is multi-variate Gaussian distribution defined over a P-dimensional vector of continuous variables:  Step 2: Estimate prior prob. of class k,  Step 3: Compute posterior prob. (the probability of each class given a sample) combining conditional with priors using Bayes’ rule:  is marginal distribution obtained by summing of classes. As usual, denominator in Bayes’ theorem can be found in terms of quantities appearing in numerator, because:  Step 4: Classify using Maximum-a-Posteriori probability: | LDA is a generative model since class-conditional distributions can be used to generate samples of each classes.  LDA is useful to deal with imbalanced group sizes (e.g.: ) since priors prob. can be used to explicitly re-balance classification by setting or whatever seems relevant.  LDA can be generalised to multiclass case with 𝐾 > 2. With , LDA lead to same solution than Fisher’s linear discriminant.  **LOGISTIC REGRESSION** is called a generalized linear model, i.e. is a linear model with a link function that maps output of linear regression to posterior probability of each class using logistic sigmoid function:  Alternatively, we can say that logistic regression:  🟑Estimates probability of belonging to category using regression on predictor variables  🟑Seeks to predict probability that output variable will fall into category based on values of independent (predictor) variables.  🟑This probability is used to classify observation into category.  Logistic regression seeks to minimizes likelihood L as loss function  In two-class case, algorithms simplify considerably by coding two-classes, and via a 0/1 response . Indeed, since , log-likelihood function can be written as:  Logistic regression is a **discriminative model** since it focuses only on posterior probability of each class . It only requires estimating *𝑃* weight of *𝑤* vector. Thus, it should be favoured over LDA with many input features. In small dimension and balanced situations, it would provide similar predictions than LDA. However imbalanced group sizes cannot be explicitly controlled. It can be managed using a reweighting of input samples.  Generally used when dependent variable is binary—takes on two values, 0 or 1  **Classification using logistic regression:**  Estimate prob. p that observation belongs to category 1, , and, consequently, probability that it belongs to category 0, .  Then use , typically 0.5, with which to compare p; classify observation into 1 of two categories.  Dependent variable is called logit, which is natural logarithm of – called odds of belonging to category 1.  The form of logistic regression model is  The logit function can be solved for p: | **k-NEAREST NEIGHBOURS (K-NN) ALGORITHM**  **Basic ideas:**  🟑 Looking at ‘k’ closest labelled data points, then taking majority vote    🟑Finds records in database that have similar numerical values of set of predictor variables  🟑Measure Euclidean distance between records in training data set. Nearest neighbour to record in training data set is 1 that that has smallest distance from it.  🟑If , then rule classifies record in same category as its nearest neighbour.  rule finds k-Nearest Neighbours in training data set to each record we want to classify; then assigns classification as classification of majority of k nearest neighbours.  🟑Typically, various values of k are used; then results inspected to determine which is best.  **Measuring model performance:**  Accuracy – Fraction of correct predictions  Could compute accuracy on data used to fit classifiers: Not indicative of ability of generalise  Split data into training and test set: Fit/train classifier on training set; make predictions on test set; compare predictions with known labels.  **Model complexity:**  Small value of K means that noise will have higher influence on result 🡪 more complex model 🡪 can lead to *overfitting*  Large value of K makes it computationally expensive; defeats basic idea behind KNN (that points that are near might have similar classes) 🡪 Smoother decision boundary 🡪 less complex model  **How to choose value k?**  Selecting value of K in K-nearest neighbour is most critical problem.  Simple approach to select k is  To optimize results, we can use CROSS VALIDATION. We can test KNN algorithm with different values of K. Model which gives good accuracy can be optimal choice.  **Model complexity and over/under fitting:**    **OVERFITTING** VC dimension (for dimension) is a **measure of capacity** (complexity, expressive power, richness, or flexibility) of a statistical classification algorithm, defined as cardinality of largest set of points that algorithm can shatter.  **Theorem**: Linear classifier in have VC dimension of 𝑃 + 1. Hence in dimension two () any random partition of 3 points can be learned | **Metrics of classification performance evaluation**  **METRICS for binary classification source:**  Imagine a study evaluating a new test that screens people for a disease. Each person taking test either has or does not have disease. Test outcome can be positive (classifying person as having disease) or negative (classifying person as not having disease). Test results for each subject may or may not match subject’s actual status. In that setting:    **True positive ()**: Sick people correctly identified as sick  **False positive ()**: Healthy people incorrectly identified as sick  **True negative ()**: Healthy people correctly identified as healthy  **False negative ()**: Sick people incorrectly identified as healthy  **Accuracy ()**: Fraction of correct prediction  **Sensitivity ()** or recall of positive class or true positive rate (TPR) or hit rate:  **Specificity ()** or **recall** of **negative class** or **true** **negative rate**:  **Precision** or **positive predictive value** (PPV):  **Balanced accuracy ()** is a useful performance measure is balanced accuracy which avoids inflated performance estimates on imbalanced datasets. “The balanced accuracy and its posterior distribution”). It is defined as arithmetic mean of sensitivity and specificity, or average accuracy obtained on either class:  **F1 Score** (or **F-score**) which is a weighted average of precision and recall are useful to deal with imbalanced datasets  The four outcomes can be formulated in a 2×2 contingency table or confusion matrix  Reference: <http://scikit-learn.org/stable/modules/model_evaluation.html>  **Significance of classification rate**  P-value associated to classification rate. Compared number of correct classifications to null hypothesis of Binomial distribution of parameters (typically 50% of chance level) and (Number of observations).  Is 60% of accuracy a significant prediction rate among 62 observations?  Since this is an exact, **two-sided** test of null hypothesis, p-value can be divided by 2 since we test that accuracy is superior to chance level. |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Area Under Curve (AUC) of Receiver operating characteristic (ROC)**  Some classifier may have found a good discriminative projection . However, if threshold to decide final predicted class is poorly adjusted, performances will highlight a high specificity and a low sensitivity or contrary.  In this case it is recommended to use AUC of a ROC analysis which basically provide a measure of overlap of two classes when points are projected on discriminative axis.  **IMBALANCED CLASSES**  Learning with discriminative (logistic regression, SVM) methods is generally based on minimizing misclassification of training samples, which may be unsuitable for imbalanced datasets where recognition might be biased in favour of most numerous class. This problem can be addressed with a generative approach, which typically requires more parameters to be determined leading to reduced performances in high dimension.  Dealing with imbalanced class may be addressed by three main ways, resampling, reweighting and one class learning.  In **sampling strategies**, either minority class is oversampled, or majority class is under sampled or some combination of two is deployed. Under sampling, majority class would lead to a poor usage of left-out samples. Sometimes one cannot afford such strategy since we are also facing a small sample size problem even for majority class.  Informed oversampling, which goes beyond a trivial duplication of minority class samples, requires estimation of class conditional distributions in order to generate synthetic samples.  Here generative models are required. An alternative proposed in generate samples along line segments joining any/all of k minority class nearest neighbours. Such procedure blindly generalizes minority area without regard to majority class, which may be particularly problematic with high-dimensional and potentially skewed class distribution.  **Reweighting**, also called cost-sensitive learning, works at an algorithmic level by adjusting costs of various classes to counter class imbalance. Such reweighting can be implemented within SVM or logistic regression classifiers. Most classifiers of Scikit learn offer such reweighting possibilities.  The parameter can be positioned into "balanced" mode which uses values of 𝑦 to automatically adjust weights inversely proportional to class frequencies in input data as . | **ASSOCIATION RULE MINING**, often called affinity analysis, seeks to uncover associations and/or correlation relationships in large data sets.  Association rules identify attributes that occur together frequently in given data set.  Market basket analysis, for example, is used determine groups of items consumers tend to purchase together.  *Association rules provide information in form of if then (antecedent consequent) statements*  **MEASURING STRENGTH OF ASSOCIATIONS:**  **SUPPORT** for (association) rule is percentage (or number) of transactions that include all items both antecedent; consequent.  **CONFIDENCE** of (association) rule is ratio of number of transactions that include all items in consequent as well as antecedent (namely, support) to number of transactions that include all items in antecedent  **EXPECTED** confidence is number of transactions that include consequent divided by total number of transactions.  **LIFT** is ratio of confidence to expected confidence.  Higher lift ratio, stronger association rule; value greater than 1.0 is usually good minimum.  Example: supermarket database has 100,000 point-of-sale transactions; 2000 include both; B items; 5000 include C;; 800 include A, B; C  Association rule: “If; B are purchased, then C is also purchased.”  Support = 800/100,000 = 0.008  Confidence = 800/2000 = 0.40  Expected confidence = 5000/100000 = 0.05  Lift = 0.40/0.05 = 8  The lift ratio indicates how much more likely we are to encounter event; B are purchased, as compared to entire population of transactions.  **Cause; Effect modelling:**  Correlation analysis can help us develop cause-and effect models that relate lagging; leading measures.  Lagging measures tell us what often external business results such has happened; are as profit, market share, or customer satisfaction.  Leading measures predict what will happen; are usually internal metrics i.e. employee satisfaction, productivity; turnover. |  |  |  |  |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Nonlinear learning algorithms  **SUPPORT VECTOR MACHINES (SVM)**  SVM are based kernel methods require only a user specified kernel function , i.e., a **similarity function** over pairs of data points into kernel (dual) space on which learning algorithms operate linearly, i.e. every operation on points is a linear combination of .  **Outline of SVM algorithm:**  🟑Map points into kernel space using a kernel function: .  🟑Learning algorithms operate linearly by dot product into high-kernel space .  ⎯ Using kernel trick (Mercer’s Theorem) replace dot product in high dimensional space by a simpler operation such that . Thus, we only need to compute a similarity measure for each pair of point and store in a Gram matrix.  ⎯ Finally, learning process consist of estimating of decision function that maximises hinge loss (of ) plus some penalty when applied on all training points.  🟑 Predict a new point using then decision function.  **Gaussian kernel (RBF, Radial Basis Function):**  One of most commonly used kernels is Radial Basis Function (RBF) Kernel. For a pair of points RBF kernel is defined as:  Where defines kernel width parameter.  Basically, we consider a Gaussian function centred on each training sample . it has a ready interpretation as a similarity measure as it decreases with squared Euclidean distance between two feature vectors.  **RANDOM FOREST**  A random forest is a meta estimator that fits several **decision tree learners** on various sub-samples of dataset and use averaging to improve predictive accuracy and control over-fitting.  **DECISION TREE LEARNER**  A tree can be “learned” by splitting training dataset into subsets based on an features value test.  Each internal node represents a “test” on a feature resulting on split of current sample.  At each step algorithm selects feature and a cut-off value that maximises a given metric.  Different metrics exist for regression tree (target is continuous) or classification tree (the target is qualitative).  This process is repeated on each derived subset in a recursive manner called recursive partitioning. Recursion is completed when subset at a node has all same value of target variable, or when splitting no longer adds value to predictions. This general principle is implemented by many recursive partitioning tree algorithms. Decision trees are simple to understand and interpret however they tend to overfit data.  However, decision trees tend to overfit training set. Leo Breiman propose random forest to deal with this issue. | Resampling method  **LEFT OUT SAMPLE’S VALIDATION**  The **training error** can be easily calculated by applying statistical learning method to observations used in its training. But because of overfitting, training error rate can dramatically underestimate error that would be obtained on new samples.  The **test error** is average error that results from a learning method to predict response on a new sample that is, on samples that were not used in training method. Given a data set, use of a learning method is warranted if it results in a low-test error. test error can be easily calculated if a designated test set is available. Unfortunately, this is usually not case.  Thus, original dataset is generally split in a training and a test (or validation) data sets. Large training set (80%) small test set (20%) might provide a poor estimation of predictive performances. On contrary, large test set and small training set might produce a poorly estimated learner. This is why, on situation where we cannot afford such split, it recommended to use cross-validation scheme to estimate predictive power of a learning algorithm  **CROSS-VALIDATION (CV)**  **MOTIVATION:**  Model performance is dependent on way data is split.  Not representative of model’s ability to generalize  **🡪 solution is Cross-validation**    **Model performance:**  5-folds = 5-fold Cross validation  🡪 k folds = k-fold Cross validation  🡪 More folds = More computationally expensive  Cross-Validation scheme randomly divides set of observations into 𝐾 groups, or **folds**, of approximately equal size. first fold is treated as a validation set, and method is fitted on remaining union of 𝐾 – 1-fold: . measure of performance (the score function ), either an error measure or a correct prediction measure is an average of a loss error or correct prediction measure, noted , between a true target value and predicted target value. score function is evaluated of on observations in held-out fold. For each sample 𝑖 we consider model estimated on data set without group that contains noted . This procedure is repeated times; each time, a different group of observations is treated as a test set. Then we compare predicted value with true value using an Error or Loss function . We can compute a score averaging over all samples:  Similarly, we can compute score S on each fold and average across folds:  these two measures (an average of average vs. a global average) are generally similar. They may differ slightly is folds are of different sizes.  This validation scheme is known as **K-Fold CV**. Typical choices of are 5 or 10, [ 1995]. extreme case where is known as **leave-one-out cross-validation, LOO-CV**. | **CV FOR REGRESSION**  Usually error function is r-squared score. However other function could be used  **CV FOR CLASSIFICATION**  With classification problems it is essential to sample folds where each set contains approximately same percentage of samples of each target class as complete set. This is called **stratification**. In this case, we will use with is a variation of k-fold which returns stratified folds.  Usually error function are, at least, sensitivity and specificity. However other function could be used.  **CV for model selection: setting hyper parameters** It is important to note CV may be used for two separate goals: 1. **Model assessment**: having chosen a final model, estimating its prediction error (generalization error) on new data.  2. **Model selection**: estimating performance of different models in order to choose best one. One special case of model selection is selection model’s hyper parameters.  Indeed, remember that most of learning algorithm have a hyper parameter (typically regularization parameter) that must be set. Generally, we must address two problems simultaneously. usual approach for both problems is to randomly divide dataset into three parts: a training set, a validation set, and a test set.  • **training set** () is used to fit models;  • **validation set** () is used to estimate prediction error for model selection or to determine hyper parameters over a grid of possible values.  • **test set** (test) is used for assessment of generalization error of final chosen model.  **GRID SEARCH PROCEDURE**  Model selection of best hyper parameters over a grid of possible values  For each possible values of hyper parameters *𝛼𝑘*:  1. Fit learner on training set:  2. Evaluate model on validation set and keep parameter(s) that minimises error measure  3. Refit learner on all training + validation data using best hyper parameters:  4. \*\* Model assessment \*\* of on test set:  **NESTED CV for MODEL SELECTION and ASSESSMENT**  Most of time, we cannot afford such three-way split. Thus, again we will use CV, but in this case, we need two nested CVs.  One **outer CV loop, for model assessment**. This CV performs splits of dataset into training plus validation set and a test set  One **inner CV loop, for model selection**. For each run of outer loop, inner loop loop performs *𝐿* splits of dataset into training setand a validation set: .  **IMPLEMENTATION WITH SCIKIT-LEARN**  Note that inner CV loop combined with learner form a new learner with an automatic model (parameter) selection procedure. This new learner can be easily constructed using . learned is wrapped inside a class.  Then new learned can be plugged into classical outer CV loop  **Regression models with built-in cross-validation**  will automatically select a grid of parameters, most of time use defaults values.  is number of CPUs to use during cross validation. If , use all CPUs. | **RANDOM PERMUTATIONS**  A permutation test is a type of non-parametric randomization test in which null distribution of a test statistic is estimated by randomly permuting observations.  Permutation tests are highly attractive because they make no assumptions other than that observations are independent and identically distributed under null hypothesis.  1. Compute an observed statistic on data.  2. Use randomization to compute distribution of *𝑡* under null hypothesis: Perform random permutation of data. For each sample of permuted data, data compute statistic . This procedure provides distribution of under null hypothesis  3. Compute p-value = where ‘s include  **BOOTSTRAPPING**  is a random sampling with replacement strategy which provides a nonparametric method to assess variability of performances scores such standard errors or confidence intervals.  A great advantage of bootstrap is its simplicity. It is a straightforward way to derive estimates of standard errors and confidence intervals for complex estimators of complex parameters of distribution, such as percentile points, proportions, odds ratio, and correlation coefficients.  1. Perform sampling, with replacement, of dataset.  2. For each sample fit model and compute scores. 3. Assess standard errors and confidence intervals of scores using scores obtained on resampled dataset  **Resampling:**  Statistical methods over different time intervals: …  DOWN SAMPLING: reduce datetime rows to slower frequency  UP SAMPLING: increase datetime rows to faster frequency |  |  |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Monte Carlo  **MONTE CARLO SIMULATION:** isprocess of generating random values for uncertain inputs in model, computing output variables of interest; repeating this process for many trials to understand dist. of output.  **PERFORM FOLLOWING STEPS:**  1. Develop visual model  2. Determine probability dist. that describes uncertain inputs in model  3. Identify output variables you wish to predict  4. Set number of trials or repetitions for simulations  5. Run simulation  6. Interpret results  **MARKET BASKET ANALYSIS**, for example, is used determine groups of items consumers tend to purchase together.  Association rules provide information in form of if then (antecedent consequent) statements**.** In other situations, historical data are not available; we can draw upon properties of common prob. dist. to help choose representative dist. that has shape that would most reasonably represent analyst’s understanding about uncertain variable.  Uniform or triangular dist. are often used in absence of data.  **SAMPLING METHODS:**  Monte Carlo sampling selects random variates independently over entire range of possible values of distribution. Monte Carlo sampling is more representative of reality; should be used if you are interested in evaluating model performance under various what-if scenarios.  **CONFIDENCE INTERVAL FOR MEAN:**  Each time you run simulation; you will obtain slightly different results.  Confidence interval:  Because Monte Carlo simulation will generally have very large number of trials, we may use standard normal z value instead of t-dist. in confidence interval formula.  **FLAWS OF AVERAGES:** evaluation of model output using average value of input is not necessarily equal to average value of outputs when evaluated with each of input values.  In newsvendor example, quantity sold is limited to smaller of demand; purchase quantity, so even when demand exceeds purchase quantity, profit is limited.  Using average values in models can conceal risk.  **MONTE CARLO USING SIMULATION USING FITTED DISTRIBUTION:**  Sampling from empirical data has some drawbacks.  Empirical data may not adequately represent true underlying population because of sampling error.  Using empirical dist. precludes sampling values outside range of actual data.  **STEPS FOR “FITTING” THEORETICAL DIST.; COMPUTING GOODNESS OF FIT:**  Choose suitable theoretical model:  For instance, normal or power law model. This task is informal; descriptive statistics like histogram; skewness indicator of observed data can be valuable hints;  Estimate model parameters:  Each theoretical model has parameters, for instance, mean; standard deviation for normal model. This task consists of estimating most likely model parameters for empirical dataset;  Determine significance level:  This tricky step establishes how good observed data match theoretical model with estimated parameters. If computed significance level is beyond pre-defined threshold, goodness-of-fit hypothesis is accepted, otherwise it is rejected  **ESTIMATE MODEL PARAMETERS:**  The maximum likelihood estimation method (MLE) is most popular method to estimate dist. parameters from empirical sample.  It finds model parameters that maximize likelihood of observed data with respect to theoretical model. | **DETERMINE SIGNIFICANCE LEVEL:**  FIT NORMAL DISTRIBUTION use Shapiro-Wilk test: If p-value is lower than threshold (usually fixed to 0.05) then normality hypothesis is rejected.  FIT ARBITRARY DISTRIBUTION use Kolmogorov-Smirnov test: If p-value is lower than given threshold, goodness-of-fit hypothesis is rejected.  **CASH-BUDGET MODEL** is process of projecting; summarizing company’s cash inflows; outflows expected during planning horizon.  Most cash budgets are based on sales forecasts. Because of inherent uncertainty in sales forecasts, Monte Carlo simulation is appropriate tool for modelling cash budgets. | Linear integer optimisation models  **BUILDING LINEAR OPTIMIZATION MODELS:**  Step 1. Identify decision variables – unknown values that model seeks to determine.  Step 2. Identify objective function – quantity we seek to minimize or maximize.  Step 3. Identify all appropriate constraints – limitations, requirements, or other restrictions that are imposed on any solution, either from practical or technological considerations or by management policy.  Step 4. Write objective function; constraints as math expressions  **LINEAR OPTIMIZATION MODEL** (often called **linear program**/**LP**) has 2 basic properties.  1. objective function; all constraints are linear functions of decision variables: This means that each function is simply sum of terms, each of which is some constant multiplied by decision variable.  2. All variables are continuous: This means that they may assume any real value (typically, nonnegative).  **HOW SIMPLEX METHOD WORKS?** simplex method evaluates impact of constraints in terms of their contribution to objective function for each variable. For simple case of only 1 constraint, optimal (maximum) solution is found by simply choosing variable with highest ratio of objective coefficient to constraint coefficient.  **Example 3:**  Manufacturing produces 4 types of structural support fittings. Machining centres have capacity of 280,000 minutes per year. Gross margin/unit; machining:   |  |  |  |  |  | | --- | --- | --- | --- | --- | | Product | Plugs | Rails | Rivets | Clips | | Gross margin/unit | 0.3 | 1.3 | 0.75 | 1.2 | | Minute/unit | 1 | 2.5 | 1.5 | 2 |   How many units of each product type should produce to maximize gross profit margin?  Objective: Maximize gross profit margin  Constraints:    Clips have highest marginal profit per unit of resource consumed.  Maximum possible production of clips  = 280,000 minutes ÷ minutes/unit  = 280,000 ÷ 2 = 140,000  Profit for maximum production of clips  = gross margin/unit \* max possible production  = $1.20 \* 140,000 = $168,000  **OUTCOMES:**  Unique optimal solution: there is exactly 1 solution that will result in maximum (or minimum) objective.  Alternative (multiple) optimal solution: objective is maximized (or minimized) by more than 1 combination of decision variables, all of which have same objective function value.  Unbounded solution: objective can be increased or decreased without bound (i.e., to infinity for maximization problem or negative infinity for minimization problem)  Infeasibility: no feasible solution exists  **SENSITIVITY ANALYSIS for DECISION variable:**  Sensitivity Analysis allows us to understand how optimal objective value; optimal decision variables are affected by changes in objective function coefficients, impact of forced changes in certain decision variables, or impact of changes in constraint resource limitations or requirements.  Sensitivity Analysis applies to changes in only 1 of model parameters at time; all others are assumed to remain at their original values  Reduced Cost: How much objective function coefficient needs to be reduced for nonnegative variable that is zero in optimal solution to become positive. | If variable is positive in optimal solution, its reduced cost is zero. If objective coefficient of any 1 variable that has positive value in current solution changes but stays within range specified by Allowable Increase; Allowable Decrease, optimal decision variables will stay same; however, objective function value will change.  **SENSITIVITY ANALYSIS for CONSTRAINTS:**  **SHADOW PRICE** - how much objective function will change as right-hand side of constraint is increased by Whenever constraint has positive slack, shadow price is zero. When constraint involves limited resource, shadow price represents economic value of having additional unit of that resource.  **USING SENSITIVITY ANALYSIS:**  If change in objective function coefficient remains within Allowable Increase; Allowable Decrease ranges, then optimal values of decision variables will not change. However, you must recalculate value of objective function using new value of coefficient.  If change in objective function coefficient exceeds Allowable Increase or Allowable Decrease limits, then you must re-solve model to find new optimal values.  If change in right-hand side of constraint remains within Allowable Increase; Allowable Decrease ranges, then shadow price allows you to predict how objective function value will change 🡪 Multiply change in right-hand side (positive if increase, negative if decrease) by value of shadow price. However, you must re-solve model to find new values of decision variables.  If change in right-hand side of constraint exceeds Allowable Increase or Allowable Decrease limits, then you cannot predict how objective function value will change using shadow price 🡪 You must re-solve problem to find new solution.  **INTEGER OPTIMIZATION:**  ***Solving models vs. General Integer Variable*:**  Decision variables that we force to be integers are called general integer variables.  Algorithms for integer optimization models first solve LP relaxation (no integer restrictions imposed); gradually enforce integer restrictions using systematic searches.  ***Sensitivity analysis for Integer Optimization*:**  Because integer models are discontinuous by their very nature, sensitivity information cannot be generated in same manner as for linear models  To investigate changes in model parameters, it is necessary to re-solve model.  **Example 1:** A company makes 110-inch wide rolls of thin sheet metal; slices them in smaller rolls of 12, 15; 30 inches.  A cutting pattern is configuration of number of smaller rolls of each type that are cut from raw stock. Six different cutting patterns are used.   |  |  |  |  |  | | --- | --- | --- | --- | --- | | Size of End Item | | | | | | Pattern | 12in. | 15in. | 30in. | Scrap | | 1 | 0 | 7 | 0 | 5in. | | 2 | 0 | 1 | 3 | 5in. | | 3 | 1 | 0 | 3 | 8in. | | 4 | 9 | 0 | 0 | 2in. | | 5 | 2 | 1 | 2 | 11in. | | 6 | 7 | 1 | 0 | 11in. |   Demands for coming week are 500 12-inch rolls, 715 15-inch rolls; 630 30- inch rolls.  Problem is to develop model that will determine how many 110-inch rolls to cut into each of six patterns to meet demand; minimize scrap.  Model development: Let be number of 110-inch rolls to cut using pattern. to be whole number (general integer variable) because each roll that is cut generated different number of end items. only constraints are end-item demand, non-negativity; integer restriction | **Workforce scheduling model** is practical, yet highly complex, problem in many businesses i.e. food service, hospitals; airlines.  Typically, huge number of possible schedules exist; customer demand varies by day of week; time of day, further complicating problem of assigning workers to time slots. |  |