Artificial Intelligence 2 (MOD Pelillo)
Grid List

Ca' Foscari University of Venice
Department of Environmental Sciences, Informatics and Statistics

Academic Year 2021 - 222

Student Zuliani Riccardo 875532

Contents

Chapter 1

Information Theory

- **Communication System**
  - Source
  - Transmitter
  - Channel
  - Noise
  - Receiver
  - Destination
- Efficiency
- Reliability
- **Redundancy**
- **Three levels of information**
  - Syntatical Level
  - Semantical Level
  - Pragmatic Level
- **Quantify Information**
  - Definition
  - $P(E) = 1$
  - $P(E) = 0$
  - Information as Probability of Function
  - Proprieties
  - **Unique Function** that satisfy the proprieties
- **Definition of Entropy**
  - Source, Stochastic Process, Entropy
  - Problem with $p(x) = 0$

  - Proprieties of entropy: $H(x) = 0 \quad H(x) = log(n)$

- **Entropy of two random variables**
  - Input
  - *Marginal Entropy*
  - *Joint Entropy*
  - *Conditional Entropy*
  - *Chain Rule*

- **Mutual Information**
  - Definition
  - Euclidean Distance, proprieties
  - Kulback-Leibler Distance, proprieties
  - Compute the mutual information
    * Useful to compute how much information travels on the channel
    * Before
    * After
    * Proprieties

Chapter 2

Data Compression - Source Coding Theorem

- **Coding definition**
- Rules that must follows
- **4 types of code:**
  - Non-singular codes
  - Unique decodable code
  - Prefix code
  - Quantity efficiency code
    - **Length of a code or Measure of efficiency**
    - Relationship between *efficiency of a code* and *entropy given by the source*
    - Entropy lower bound for L(c), D-acid
- Huffman Coding
- Optimality of Huffman Code

Chapter 3

Reliable Communication Through Unreliable Channels

- Types of channel
  - Lossless code
  - Lossy code
- **Formal definition of Channel**
- Channel representation
  - Channel graph
  - Channel matrix
- N-th extension of the channel
- **Capacity of the channel**
  - Explanation and analysis of the *mutual information*
  - It depends on source and channel
  - **Capacity of the channel definition**
- **Reliability**
  - Code replication
  - Improve reliability
  - Reduces the rate speed of sanding data
  - Trade off
  - Rateo of speed
  - **Shannon's Second Theorem**

Chapter 4

Neural Networks

- Paradigm inspired by the way biological nervous system works, elements: neurons
- **McCulloch and Pitts Model**
- Network Topologies and Architectures
  - Feedforward only: fully connected and single layer
  - Feedback networks: sparsely connected and multilayer
- Classification Problem
  - Features, classes
  - Finding the best configuration of weights on the income connection and the threshold
  - Forget the threshold by adding an extra unit set to -1
- **Perceptron:** definition
- **Perceptron Learning Algorithm**, parameters
  1. Initialization
  2. Activation
  3. Computation of actual response
  4. Adaption of weight vector
  5. Continuation
- The perceptron Convergence Theorem
- Multilayer feedforward networks
  - Single layer
  - Multilayer feedforward network by adding a hidden layer
  - **Universal Approximation Power**
- **Back - propagation Learning Algorithm**

- Definition, Supervised Learning
- In what consist the learning
- Error Function
- What is our aim
- What do we use to achieve this
- **Pass:**
  - Feedforward Pass
  - Backword Pass
    - Notation
    - *Updating Hidden - to - Output Weights*
    - *Updating Input - to - Hidden Weights*
- Locality of Back - propagation
  - Off - Line
  - On - Line
  - Compromise

- **The Algorithm**
- Problem of the choice of the learning rate:
  - Small
  - Big
  - Solution: momentum term: definition and characteristics
- Problem of local minima
- Theoretical / Practical questions
  - Generalization
  - Training, Validation and Test set
  - Learning phase stopped in the minimum validation error
- Model Evaluation
  - True error
  - Sample error
- Cross validation
- Overfitting
- Size of Neural Network -¿ Horizon Effect
- **Pruning Approach:**
  - Definition
  - Online and Offline Pruning

- What we have to do
- Algebra and Vector description
- Consider only the initial contributions, with the heuristic that it will not far away from the real one
- Algorithm

Chapter 5

Optimal Brain Surgeon

- Usage of the second order derivatives to improve generalization
- Permits pruning of more weights than other methods
- **Key point** is the recursion relation for calculating the inverse Hessian matrix H^{-1}
- **Introduction**
  - Problem: minimize the system complexity
  - Casted in minimizing the number of connection weights
  - This overfitting could occur
  - Which weight should be eliminate?
  - Delete weights with small magnitude, but lead to wrong weights
  - OBD uses the minimal increase in training error for weight elimination
  - Assume that the matrix is diagonal
  - ODB delete wrong weights
  - OBS makes no restrictive assumption about the form of the Hessian
- **Optimal Brain Surgeon**
  - Function of Taylor series respect to weights
  - etc

Chapter 6

Statistical Learning Theory

- Deal with supervised learning (input(feature), output(label))
- Estimate a function relationship between the input and the output spaces
- **Classification Algorithm**
- Assumption
  - Joint probability distribution unknown
  - learning example sampled independently
  - No assumption on p is made
  - p is fixed
- Measure of "How good" a function f is when we use a classifier
- Loss function
- Risk
- Best classifier
- The classification problem
- Nearest Neighbour Classifier
  - Definition
  - Assumption: training set is fixed
  - K-NN
  - $K_n - NN$
  - How good is the Nearest Neighbour rule
  - Stone Theorem
  - Kernel rule, smoothing factor
- **Empirical Risk Minimization Principle**
  - Minimize empirical risk
    - Training data

    * Family of function
    * Loss function
    * Empirical Risk Minimization (ERM)
- Estimation VS Approximation
- Small complexity on F
- Large complexity on F
- Shattering
- VC Dimension
- Structural Risk Minimization

Chapter 7

Deep Neural Network

- Learn feature hierarchy from the initial pixel in order to obtain a classifier
- Shallow Architecture
  - Inefficient to represent deep features
  - Universal Approximation Law
  - Produces large hidden layer and increase a lot the number of parameters
- Deep Architecture
  - Fit function better with less parameter
  - Increase the number of hidden layer
  - Decrease the number of parameter
- Idea not new
- More available data, more computing power, new idea
- Image classification, image, height x weight x 3 of 0 x 255 dimension
- Challenge
  - Viewpoint
  - Illumination
  - Scale
  - Deformation
  - Background clutter
  - Occlusion
  - Intraclass
- Data driven approch
- Retina, Relative field, feature detector
- Cat experiment
- Specialized Neuron, activating by lines, edges etc

- Convolution
- Matrix dot product with filter
- **Convolution**
- Mask: identity, edge detection, sharpeon, box blur, gaussian blur
- Strade, Padding
- Traditional approach and deep learning
- **Convolutional Neural Networks (CNN)**
  - Object Character Recognition
  - Fully Connected NN
  - Locally Connected NN
- Maxpooling
- **AlexNET**
  - 8 layer in the following schema:
    1. Conv + Pool
    2. Conv + Pool
    3. Conv
    4. Conv
    5. Conv + Pool
    6. Full
    7. Full
    8. SoftMax Output (1000 way)
  - 2 independent GPU, run in parallel and there are connection between the two GPUs
  - Way GPU
  - Deepening in softmax
  - Using the sigmoid activation function to propagate G, becomes zero
  - ReLU
- Mini-Batch stochastic gradient descent
- Technique to reduce overfitting
  - Data Argumentation
  - Dropout
- **ImageNET**
  - What is and how it has been build
  - Overall

- Other Computer vision task
  - Semantic segmentation
  - Classification + Localization
  - Object Detection
  - Instance Segmentation
  - Image Captioning
- **Recurrent Neural Network**

Chapter 8

Support Vector Machine (SVM)

- Abstract idea of SVM
- Formal definition of SVM classifier $h_{w,b}(x) \quad g(z)$
- Confidence
- Example of SVM
- Example of many decision boundary
- Question
- **Functional margin**
- **Geometric margin**
- Relation between the two margins
- The three steps of the optimization problem, optimal margin classifier
- **Duality**
  - Actual Problem
  - Lagrangian function / Duality
  - Lagrangian Dual Function
  - Lower bound on optimal value
  - Focus on convex problem
  - Wolfe Duality
- Primal Problem
- N lagrangian, so derivatives respect to w and b, with merging in the lagrangian
- Our dual problem, rephrasing of the optimization problem
- Dot product, only support vectors are used
- Maximum margin hyperplane

- Given the solution of the dual optimization probelm
  - Weight vector of the maximum margin hyperplane
  - Hyperplane
  - Linear SVM classifier
- Support Vectors Equation
- SVM error function
- SVM and VC dimension (Vapnik Theorem)
- Outliers of soft margin
  - Correct classify it reducing the roboustness
  - Leave it missclassified, reducing the effect by introducing slack variable
  - Formula
  - Small C
  - Large C
  - C ti infinity
  - Dual representation of the problem
  - Soft margin as hard margin
  - Formula
- Kernel Trick
  - SVM works with linearly separable problem
  - Kernel Trick to learn a hyperplane in a new space, interesting when data is not linearly separable
  - Mapping function
  - Kernel function
  - Cover's Theorem
  - SVM works with inner product between input vector
  - Replace with a kernel function to learn in a new feature space
  - Restrictions:
    * Marcer's Theorem
    * Positive definite kernel
  - The discriminant function / hyperplane
  - Replacing the map function in the dual problem
- **Multi - Class Problems**
  - One-vs-the rest classifier
  - One-vs-one classifier
  - best approach: k one-vs-one classifier and use the most accurate classifier

Chapter 9

Unsupervised Learning

- Classical clustering problem: set of n objects, n x n matrix A of pairwise similarities
- Goal is to partition the vertices of G into maximally homogeneous groups (clusters)
- **K-Means**
  - Description
  - How it works
  - Proprieties
  - Problem
- **Image as a graph**
  - Description
  - Feature Base (Central) Clustering
  - Graph base (Pairwise) Clustering
  - Gaussian Kernel
- **Eigenvector - Based Clustering**
  - Cluster as a vector x (participation of each node)
  - Normalize the eigenvectors
  - We want to maximize
  - Eigenvalue problem, chose the eigenvector of A with largest eigenvalue
  - If $A = A^T$ then A is symmetric and has only real eigenvalues
  - If A is symmetric then $x^T Ax$
  - **The algorithm:**
    1. Affinity matrix A
    2. Eigenvalues and eigenvectors
    3. Repeat
       (a) Eigenvector of the larges unprocessed eigenvalue

   (b) Zero components that has been processed
   (c) Threshold the other to determine its belonging
   (d) All elements processed, there are suff clusters
4. Until there are suff clusters

- Clustering as graph partitioning
  - Formula
  - Minimum cut problem
  - Advantage: poly time
  - Disadvantage: measure what happens between the two clusters and not within both
- Normalize NCut
- **Graph Laplacian**
  - Main tool for spectrul clustering, unormalized graph laplacian
  - D diagonal matrix
  - W affinity matrix
  - Proprieties of matrix L:
    * $x^T x = \frac{1}{2} \sum_{i,j=1} w_{i,j}(x_i - x_j)^2$
    * L is symmetric
    * Smallest eigenvalue is 0 with eigenvector 1
    * L has non -negative eigenvalues
  - **The Normalized Graph Laplacian**
    * Symmetric Matrix
    * Random Walk Matrix
- **Solving NCut**
  - Any cut can be represented as a binary indicator vector x
  - Formula
  - y is an indicator vector
  - NP hard problem
  - Approximation
  - Relaxation of the constraint value from being discrete to continuous real value
  - Generalized eigenvalue problem
- **Two-way NCut**
  1. Affinity matrix W and degree matrix D
  2. Solve the generalized eigenvalue problem

3. Use the eigenvector associated to the second smallest eigenvalue to bipartite the graph. Way the second smallest?

   - Through relaxation we loose some precision / information. Not guaranteed that there is a one-to-one correspondance
   - Some point could not so clear to assign
   - No clear threshold to split based on the second vector
   - Shortcut to overcome this problem
     * Constant value
     * Median value
     * Splitting point that has the minimum NCut value (choose n possible splitting point, compute the NCut and choose the minimum)
   - What if we consider more than two cluster?

- **NCut with more than two clusters**
  1. **Recursive two-way NCut:**
     (a) Given G compute D and W
     (b) Solve the generalized eigenvalue problem for the smallest eigenvalue
     (c) second eigenvalue, eigenvector, bipartite the graph by finding the splitting point that minimize ncut
     (d) Decide if the current partition is satisfied or not
     (e) Continue the repartition
     - Use only the second eigenvalue
  2. **Using the first K eigenvectors:**
     (a) Unnormalized graph laplacian
     (b) K smallest eigenvectors of the generalized eigenproblem
     (c) U = u1, u2,...uk
     (d) Yi vector corresponding to the i-th row of U
     (e) Yi as points, cluster them using kmeans
- **Spectra Clustering Vs K-Means**
  - Cluster data that is connected but not necessary compact
  - Given: similarity matrix S and k number of cluster
    1. Similarity graph and normalized graph laplacian L sym
    2. lower dimension space where clusters are more obvious
    3. V = v1, v2,..., vk
    4. matrix U from V by normalizing the row sum to have norm 1
    5. Yi vector corresponding to the i-th row of U
    6. Cluster the points Yi using k-means
- K-means to laplacian eig. cluster with non convex boundaries
- Problem: choosing k s.t. all eigenvalues are very small and the next is very large
- Eigengap heuristic (difference between consecutive eigenvalue)

Chapter 10

Dominant Set

- Data rep. as weight graph so construct similarity matrix
- Data as nodes
- Edges as similarity relation between nodes
- Allow to codify and use complex structured and unstructured data
- **Cluster** maximal clique of a graph
  - Clique related to internal cluster criteria
  - Maximal clique problem can not be applied on weight graph
  - Dominant set, evolution of the maximal clique problem
- **Dominant Set**
  - measure of cohesiveness of a cluster and vertex participation of diff cluster
  - graph theory, game theory and QOP
  - Connection between Dominant set and local extrema of QOP
  - Consider nodes belonging to diff cluster considering the hp of overlapping clusters
- **Graph-theoretic definition of a cluster**
  - Data = G (V, E, w)
  - G as an adjacency matrix A
  - High *Internal* Homogeneity
  - High *External* In-Homogeneity
  - Idea of the criterion
  - S in or equal to C and i in S
  - Average weight degree of i with regard the set S
  - Relative similarity bet i and j respect to the average similarity bet i and its neighbours
  - Weigh of i with regard to S, gives the similarity bet i and S - i with respect to the overall similarity among the vertices of S - i

  - The total weight of S
  - Definition of Dominant Set
    * Internal homogeneity: all node in the cluster are important for it
    * External homogeneity: considering a new point to add the cluster cohesiveness will decrease

- **From dominant set to local optima:**
  - Vector as participation of the nodes
  - Eigenvalue problem, with A symmetric matrix
  - Problem finding x that minimize f, but has to be normalized, constraint, probability space
  - **Support of x**
  - **Characteristic vector**
  - Dominant set one-to-one correspondence with strict local maxima of quadratic function

- **Link to Game Theory**
  - Definition
  - Proprieties:
    * Symmetric game
    * Complete knowledge
    * Non-cooperative game
    * Pre-existing set of pure strategies
  - V pure strategies
  - Similarity matrix A represent the payoff matrix and it resume the revenue
  - Mixed strategy: prob dist over the set of pure strategies
  - Expected payoff of couple of player playing diff strategy
  - Goal: maximise the its resulting revenue
  - A as similarity matrix so players to maximize their revenue has to coordinate their strategy so the sampled one belongs to the same cluster
  - The players reach the symmetric mash equilibrium
  - **Nash Equilibrium**
    * Definition
    * Inequality of definition
    * Equilibrium is symmetric when x1 = x2, inequality
    * Pro: sat int hom
    * Con: not include any constraint that guarantees the maximality conditions
  - **Evolutionary Stable Strategy**
    * Definition

    * Inequalities
    * Play x since the payoff against itself is higher than y
  - **In conclusion:**
    * Clustering game one-to-one dominant set
    * Dominant set one-to-one local solution of SQOP
    * **EESs one-to-one to local solution of SQOP**
  - EES abstract well the definition of cluster
    * Internal coherency: high support for elem within the cl
    * External coherency: low support for elem out of the group
  - EES one-to-one Maximal clique, definition of clique and maximal clique

- **Extracting Dominant Set: Replicator Dynamics**
  - Individuals are repeatedly sampled at random, infinite population, to play a two-player game
  - Not suppose to have complete knowledge on the game
  - They act:
    * According to inherited behavioural
    * Pure Strategy
  - Suppose to have some Evolutionary Selection Process that operates over time
  - $x_i(t)$ population share playing pure strategy i at time t
  - Stochastic process of state of pop at time t
  - Evolution equation taken by the Darwin's principle of nature selection
  - description of it
  - Proportionality
  - Replication equation used by replicator dynamics
    * Formula
    * x_i proportion of strategy i in the pop
    * $x = (x_1, ..., x_n)$ vector of dist of strategy
    * $f_i(x)$ fitness of strategy i
    * $o_(x)$ average pop fitness
    * $\dot{x}_i$ grow rate of strategy i, it increase if
  - $f_i(x)$ is assumed to depend linearly upon the population distribution
  - Formula
  - $(Ax)_i$ expected payoff of the i-th row
  - $x^A x$ is the average payoff
  - Discretization which assume *non-overlapping* generations, formula