Data mining

Giacomo Fantoni
Telegram: @GiacomoFantoni
Github: https://github.com/giacThePhantom/DataMining

December 20, 2021

Contents

Chapter 1

Introduction

A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T as measured by P, improves with experience E.
From that it is understood that machine learning is a form of inductive learning: it generalize from examples to a concept. There is no certainty of correctness.

1.1 Formalization of a machine learning problem

1.1.1 Components of a machine learning problem

The components of a machine learning problem are:

- Task to be addressed by the system.
- Performance measure to evaluate the learned system.
- Training experience to train the learning system.

1.1.2 Designing a machine learning system

The designing of a machine learning system can be described in a process that consist of different phases:

1. Formalize the learning task.
2. Collect data.
3. Extract features.
4. Choose class of learning models.
5. Train model.
6. Evaluate model.

1.1.2.1 Formalize the learning task

To formalize the learning task means to define the task that should be addressed by learning system. This type of task, or learning problem, is often composed of a number of related tasks, sub-problems or side-tasks. It is also needed an appropriate performance measure for evaluating the learned system.

1.1.2.2 Collect data

To collect the data means to collect a set of training example in machine readable format. Data collection is often the most cumbersome part of the process, implying manual intervention especially in labelling examples for supervised learning. Recent approaches to the problem of data labelling try to make use of the availability of unlabelled data (semi-supervised learning it tries to learn using both supervised and unsupervised examples).

1.1.2.3 Extract features

A relevant set of features need to be extracted from the data in order to provide inputs to the learning system. Prior knowledge is usually necessary in order to choose the appropriate features for the task in mind. Extracting too few features can miss relevant information preventing the system from learning the task with reasonable performance. Extracting too many feature like can make the learning problem harder and require a number of examples greater than those available for training. Another problem arises when considering noisy features. It is noticeable that there is a need to choose the correct number and type of feature to permit a correct and efficient solution.

1.1.2.4 Choose class of learning models

Every problem has a class of learning model that is able to learn it best. A simple model like a linear classifier is easy to train but insufficient for non linearly separable data. A too complex model can memorize noise in training data failing to generalize to new examples. The algorithm need not to optimize but it needs to generalize, there can be outliers or labelling error. The more complex the model the more it will tend to overfit training noise.

1.1.2.5 Train model

Training a model implies searching though the space of possible models given the chosen model class. Such search typically aims at fitting the available training examples well according to the chosen performance measure. However the learned model should perform well on unseen data (generalization) and not simply memorize training examples (overfitting). Different techniques can be used to improve generalization, usually by trading off model complexity with training set fitting.

1.1.2.6 Evaluate model

The learned model is evaluated according to its ability to generalize to unseen examples. These example are collected in a test set. Evaluation can provide insights into the model weaknesses and suggest directions for refining and modifying it. Evaluation can imply comparing different models or learners in order to decide the best performing one. Statistical significance of observed differences between performance of different models should be assessed with appropriate statistical tests.

1.2 Learning settings

1.2.1 Supervised learning

The learner is provided with a set of inputs/output pairs $(x_i, y_i) \in X \times Y$. The learned model $f : x \to Y$ should map input examples into their output. A domain expert is typically involved in labelling input examples with output examples in the training set.

1.2.1.1 Tasks

1.2.1.1.1 Classification

- Binary: assign an example to one of two possible classes often a positive and a negative one.
- Multiclass: assign an example to one of $n > 2$ possible classes.
- Multilabel: assign an example to a subset $m \leq n$ of the possible classes.

1.2.1.1.2 Regression Assign a real value to an example.

1.2.1.1.3 Ordinal regression or ranking Order a set of examples according to their relative importance or quality with respect to the class.

1.2.2 Unsupervised learning

The learner is provided a set of input examples $x_i \in X$ with no labelling information. The learner models training examples, for examples clustering them together into clusters according to their similarity.

1.2.2.1 Tasks

1.2.2.1.1 Dimensionality reduction Reduce dimensionality of the data maintaining as much information as possible.

1.2.2.1.2 Clustering Cluster data into homogeneous groups according to their similarity.

1.2.2.1.3 Novelty detection Detect novel examples which differ from the distribution of a certain set of data.

1.2.3 Semi-supervised learning

The learner is provided with a set of input output pairs $(x_i, y_i) \in X \times Y$. A typically much bigger additional set of unlabelled examples $x_i \in X$ is also provided. Like in supervised learning the learned model $f : X \to Y$ should map input examples into their output. Unlabelled data can be exploited to improve performance, by forcing the model to produce similar outputs for similar inputs, or by allowing to learn a better representation of examples.

1.2.4 Reinforcement learning

The learner is provided a set of possible states S and for each state a set of possible actions A moving it to a next state. In performing action a from state s the learner is provided an immediate reward $r(s, a)$. The task is to learn a policy allowing to choose for each state s the action a maximizing the overall reward. The learner has to deal with problems of delayed reward coming from future moves and trade-off between exploitation and exploration. Typical application include moving policies for robots and sequential scheduling problems in general.

1.3 Probabilistic reasoning

Probabilistic reasoning is the reasoning in presence of uncertainty. It evaluates the effect of a certain piece of evidence on other related variables. It estimates probabilities and relations between variables from a set of informations. They depends on variables and their relations.

1.4 Choice of learning algorithms

1.4.1 Based on information available

- Full knowledge of probability distributions of data: Bayesian decision theory.
- Form of probabilities known, parameters unknown: parameter estimation from training data.
- Form of probabilities unknown, training examples available: discriminative methods: do not model input data, learn a function predicting the desired output given the input.
- Form of probabilities unknown, training examples unavailable: unsupervised methods, cluster examples by similarity.

Chapter 2

Decision trees learning

2.1 Introduction

Decision trees tend to be interpretable, it is easy to see the reason for a certain decision. They represent a disjunction of conjunctions of constraints over attribute values. Each path from the root to a leaf is a conjunction of the constraints specified in the nodes along it: they can be seen as a disjunctive normal formula. In this way every class can be written as a DNF. The leaf contains the label to be assigned to instances reaching it. The disjunction of all paths is the logical formala expressed by the tree.

2.1.1 Appropriate problems for decision trees

The class of problems that can be solved by decision trees are:

- Binary or multiclass classification with an extension to regression (with a linear regression as the leaf).
- Instances represented as attribute-value pairs.
- Different explanations for the concept are possible (disjunction).
- Some instances have missing attributes, its dealing done with probabilistic models.
- There is need for an interpretable explanation of the output. In fact the main reason for using decision tress is interpretability.

2.2 Learning decision trees

2.2.1 Greedy top-down strategy

For each node, starting from the root with full training set:

1. Choose best attribute to be evaluated.
2. Add a child for each attribute value.
3. Split node training set into children according to value of chosen attribute.

4. Stop splitting a node if it contains examples from a single class or there are no more attributes to test.

It is also known as the divide et impera approach.

2.2.2 Choosing the best attribute

A measure to choose the attribute is entropy. It measures the amount of information contained in a collection of instances S which can take a number c of possible values:

$$H(s) = -\sum_{i=1}^{c} p_i \log_2 p_i$$

Where p_i is the fraction of S taking value i. In our case instances are training examples and values are class labels. The entropy of a set of labelled examples measures its label's inhomogeneity.

2.2.2.1 Information gain

Expected reduction in entropy obtained by partitioning a set S according to the value of a certain attribute A.

$$IG(S, A) = H(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} H(S_V)$$

Where $Values(A)$ is the set of possible values taken by A and S_v is the subset of S taking value v at attribute A. The second term represents the sum of entropies of subsets of examples obtained partitioning over A values, weighted by their respective sizes. An attribute with high information gain tends to produce homogeneous groups in terms of labels, thus favouring their classification. The idea is to use the greedy strategy in which at each choice it is chosen the attribute with the maximum information gain.

2.3 Issues in decision tree learning

2.3.1 Overfitting avoidance

Requiring that each leaf has only examples of a certain class can lead to very complex trees. A complex tree can easily overfit the training set, incorporating random regularities not representative of the full distribution, or noise in the data. It is possible to accept impure leaves, assigning them the label of the majority of their training examples. Two possible strategies to prune a decision tree are:

- Pre-pruning; decide whether to stop splitting a node even if it contains training examples with different labels.
- Post-pruning: learn a full tree and then prune it.

2.3.2 Post-pruning

In post-pruning you expand the tree and then you prune it. It is introduced the validation set.

1. For each node in the tree evaluate the performance on the validation set when removing the subtree rooted at it.

2. If all node removals worsen performance, stop.
3. Choose the node whose removal has the best performance inprovement.
4. Replace the subtree rooted at it with a leaf.
5. Assign to the leaf the majority label of all examples in the subtree.
6. Return to 1.

2.3.3 Dealing with continues valued attributes

Continuous valued attributes need to be discretized in order to be used in internal node tests. Discretization threshold can be chosen in order to maximize the attribute quality criterion.

1. Examples are sorted according to their continuous attribute values.
2. For each pair of successive examples having different labels, a candidate threshold is placed as the average of the two attribute values.
3. For each candidate threshold the infogain achieved splitting examples according to it is computed.
4. The threshold producing the higher infogain is used to discretize the attribute.

2.3.4 Alternative attribute test measures

The information gain criterion tends to prefer attributes with a large number of possible values. As an extreme the unique ID of each examples is an attribute perfectly splitting the data into singletons, but it will be no use on new examples. A measure of such spread is the entropy of the dataset with respect to the attribute value instead of the class value:

$$H_A(S) = -\sum_{v \in Values(A)} \frac{|S_V|}{|S|} \log_2 \frac{|S_V|}{|S|}$$

The gain ration measures downweights the information gain by such attribute value entropy:

$$IGR(S, A) = \frac{IG(S, A)}{H_A(S)}$$

2.3.5 Handling attributes with missing values

Assume example x with class $c(x)$ has missing value for attribute A. When attribute A is to be tested at node n:

- Simple solution: Assign to x the most common attribute values among examples in n or the most common of examples in n with class $c(x)$.
- Complex solution: Propagate x to each of the children of n with a fractional value equal to the proportion of examples with the corresponding attribute value.

The complex solution implies that at test time, for each candidate class, all fractions of the test example which reached a leaf with that class are summed, and the example is assigned the class with highest overall value.

2.4 Random forest

Random forests are an ensemble of decision trees. An ensemble is a method by which predictions are given by a set of predictors and is taken the prediction most represented. They improve stability and accuracy of the predictions. Random forests are effective and one of the methods of choice in case of tabular data.

2.4.1 Training

To train a random forest:

1. Given a training set of N examples, sample N examples with replacement.
2. Train a decision tree on the sample, selecting at each node m features at random among which to choose the best one.
3. Repeat the first two step M times in order to generate a forest of M trees.

2.4.2 Testing

To test a random forest:

1. Test the example with each tree in the forest.
2. Return the majority class among the predictions.

Chapter 3

K-nearest neighbours

3.1 Introduction

The K-nearest neighbours is an algorithm that, given a training set represented as a vector of features, gives the label for a new sample as label of the majority of the K nearest sample in the training set.

3.2 Measuring the instance between instances

3.2.1 Metric or distance definition

Given a set $\mathcal{X}$ a function $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_0^+$ is a metric for $\mathcal{X}$ if for any $x, y, z \in \mathcal{X}$ the following properties are satisfied:

- Reflexivity $d(x, y) = 0 \Leftrightarrow x = y$.
- Symmetry $d(x, y) = d(y, x)$.
- Triangle inequality $d(x, y) + d(y, z) \geq d(x, z)$.

3.2.2 Euclidean distance

The euclidean distance in $\mathbb{R}^n$ is:

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

3.3 Algorithms

3.3.1 Classification

: `Knn-Classification()`

foreach *test examples* x **do**
  foreach *training examples* (x_i, y_i) **do**
    compute distance $d(x, x_i)$
  select k-nearest neighbours of x
  %`return class of` x `as majority class among neighbours`
  return $\arg\max_{x_y} \sum_{i=1}^{k} \delta(y, y_i)$

3.3.2 Regression

: `Knn-Regression()`

foreach *test examples* x **do**
  foreach *training examples* (x_i, y_i) **do**
    compute distance $d(x, x_i)$
  select k-nearest neighbours of x
  %`return the average output value among neighbours`
  return $\frac{1}{k} \sum_{i=1}^{k} y_i$

3.4 Characteristics

- Instance-based learning: the model used for prediction is calibrated for the test example to be processed.
- Lazy learning: the computation is mostly deferred to the classification phase.
- Local learner: assumes prediction should mainly influenced by nearby instances.
- Uniform feature weighting: all feature are uniformly weighted in computing dinstances.

3.5 Distance weighted k-nearest neighbour

The distance weighted k-nearest neighbour is a variant of the classic k-nearest neighbour in which the distance is weighted. The weight of a point is calculated as:

$$w_i = \frac{1}{d(x, x_i)}$$

The class is decided for classification according to the formula:

$$\arg\max_{x_y} \sum_{i=1}^{k} w_i \delta(y, y_i)$$

For regression the formula is instead:

$$\frac{\sum_{i=1}^{k} w_i y_i}{\sum_{i=1}^{k} w_i}$$

Chapter 4

Linear algebra

4.1 Vector space

A set $\mathcal{X}$ is called a vector space over $\mathbb{R}$ if addition and scalar multiplication are defined and satisfy for all $x, y, z \in \mathcal{X}$ and $\lambda, \mu \in \mathcal{R}$:

- Addition:
  - Association: $x + (y + z) = (x + y) + z$.
  - Commutation: $x + y = y + x$.
  - There is an identity element: $\exists 0 \in \mathcal{X} : x + 0 = x$.
  - There is an inverse element: $\forall x \in \mathcal{X} \exists x' \in \mathcal{X} : x + x' = 0$.
- Scalar multiplication:
  - Is distributive over elements: $\lambda(x + y) = \lambda x + \lambda y$.
  - Is distributive over scalars: $(\lambda + \mu)x = \lambda x + \mu x$.
  - Is associative over scalars: $\lambda(\mu x) = (\lambda\mu)x$.
  - There is an identity element: $\exists 1 \in \mathbb{R} : 1x = x$.

4.1.1 Properties and operations

4.1.1.1 Subspace

A subspace is any non-empty subset of $\mathcal{X}$ being itself a vector space.

4.1.1.2 Linear combination

Given $\lambda_i \in \mathbb{R} \wedge x_i \in \mathcal{X}$, a linear combination is:

$$\sum_{i=1}^{n} \lambda_i x_i$$

4.1.1.3 Span

The span of vectors $x_1, \ldots, x_n$ is defined as the set of their linear combination:

$$\{\sum_{i=1}^{n} \lambda_i x_i, \lambda_i \in \mathbb{R}\}$$

4.1.1.4 Linear independence

A set of vector x_i is linearly independent if none of them can be written as a linear combination of the others.

4.1.2 Basis

A set of vectors x_i is a basis for $\mathcal{X}$ if any element in $\mathcal{X}$ can be uniquely written as a linear combination of vectors x_j. The vectors x_j need to be linearly independent. All bases of $\mathcal{X}$ have the same number of elements, called the dimension of the vector space.

4.2 Matrices

4.2.1 Linear maps

Given two vector spaces $\mathcal{X}$ and $\mathcal{Z}$ a function $f : \mathcal{X} \to \mathcal{Z}$ is a linear map if $\forall x, y \in \mathcal{X} \lambda \in \mathbb{R}$:

- $f(x + y) = f(x) + f(y)$.
- $f(\lambda x) = \lambda f(x)$.

4.2.2 Linear maps as matrices

A linear map between two finite dimensional spaces $\mathcal{X}$ and $\mathcal{Z}$ of dimension n and m can always be written as a matrix. Let $\{x_1, \ldots, x_n\}$ and $\{z_1, \ldots, z_m\}$ be some bases for $\mathcal{X}$ and $\mathcal{Z}$ respectively. For any $x \in \mathcal{X}$:

$$f(x) = f(\sum_{i=1}^{n} \lambda_i x_i) = \sum_{i=1}^{n} \lambda_i f(x_i)$$
$$f(x_i) = \sum_{j=1}^{m} a_{ji}^m a_{ij} z_j$$
$$f(x) = \sum_{i=1}^{n}\sum_{j=1}^{m} \lambda_i a_{ji} z_j = \sum_{j=1}^{m}\big(\sum_{i=1}^{n} \lambda_i a_{ji}\big) z_j = \sum_{j=1}^{m} \mu_j z_j$$

4.2.2.1 Matrix of basis transformation

A matrix can be used to transform the basis is:

$$M \in \mathbb{R}^{m \times n} = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \vdots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix}$$

The mapping from basis to basis coefficient is done:

$$M\lambda = \mu$$

4.2.2.2 Matrix changing the coordinates, 2D examples

Let $B = \left\{ \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix} \right\}$ be the standard basis in $\mathbb{R}^2$ and $B' = \left\{ \begin{bmatrix} 3 \\ 1 \end{bmatrix}, \begin{bmatrix} -2 \\ 1 \end{bmatrix} \right\}$ be an alternative basis. The change of coordinate matrix from B' to B is:

$$P = \begin{bmatrix} 3 & -2 \\ 1 & 1 \end{bmatrix}$$

So that:

$$[v]_B = P[v]_{B'} \qquad \wedge \qquad [v]_{B'} = P^{-1}[v]_B$$

For arbitrary B and B' P's columns must be the B' vectors written in terms of the B ones.

4.2.3 Matrix properties

4.2.3.1 Transpose

The transpose matrix is the matrix obtained exchanging the rows with column M^T.

$$(MN)^T = N^T M^T$$

4.2.3.2 Trace

The trace is the sum of the diagonal elements of a matrix:

$$tr(M) = \sum_{i=1}^{n} M_{ii}$$

4.2.3.3 Inverse

The inverse is the matrix which multiplied with the original matrix gives the identity:

$$MM^{-1} = I$$

4.2.3.4 Rank

The rank of an $n \times m$ matrix is the dimension of the space spanned by its columns.

4.2.4 Matrix derivatives

$$\begin{aligned}
\frac{\partial M_X}{\partial x} &= M \\
\frac{\partial y^T M x}{\partial x} &= M^T y \\
\frac{\partial x^T M x}{\partial x} &= (M^T + M)x \\
\frac{\partial x^T M x}{\partial x} &= 2Mx \qquad \text{if } M \text{ is symmetric} \\
\frac{\partial x^T x}{\partial x} &= 2x
\end{aligned}$$

Results are columns vectors. Transposing the matrix gives the row vectors.

4.2.5 Metric structure

4.2.5.1 Norm

A function $||\cdot|| : \mathcal{X} \to \mathbb{R}_0^+$ is a norm $\forall x, y \in \mathcal{X}, \lambda \in \mathbb{R}$:

- $||x + y|| \leq ||x|| + ||y||$
- $||\lambda x|| = |\lambda| ||x||$
- $||x|| > 0 if x \neq 0$

4.2.5.2 Metric

A norm defines a metric $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_0^+$:

$$d(x, y) = ||x - y||$$

Not every metric gives rise to a norm.

4.2.6 Dot product

A dot product $\langle \cdot, \cdot \rangle : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a symmetric bilinear form which is positive semi-definite:

$$\langle x, x \rangle \geq 0 \forall x \in \mathcal{X}$$

A positive definite dot product satisfies:

$$\langle x, x \rangle = 0 \Leftrightarrow x = 0$$

4.2.6.1 Norm

Any dot product defines a corresponding norm via:

$$||x|| = \sqrt{\langle x, x \rangle}$$

4.2.6.2 Properties

4.2.6.2.1 Angle The angle θ between two vectors is defined as:

$$\cos\theta = \frac{\langle x, z\rangle}{||x||||z||}$$

4.2.6.2.2 Orthogonal Two vectors are orthogonal if $\langle x, y\rangle = 0$.

4.2.6.2.3 Orthonormal A set of vectors $\{x_1, \ldots, x_n\}$ is orthonormal is

$$\langle x_i, x_j\rangle = \delta_{ij}$$

Where $\delta_{ij} = 1$ if $i = j$, 0 otherwise. If x and y are n-dimensional column vectors, their dot product is computed as:

$$\langle x, y\rangle = x^T y = \sum_{i=1}^{n} x_i y_i$$

4.3 Eigenvalues and eigenvectors

Given a $n \times n$ matrix M the real value λ and non zero vector x are eigenvalue and the corresponding eigenvector of M if:

$$Mx = \lambda x$$

4.3.1 Cardinality

An $n \times n$ matrix has n eigenvalues, but less than n distinct ones. The number of eigenvalues is the number of linear independent eigenvectors.

4.3.2 Singular matrices

A matrix is singular if it has a zero eigenvalue:

$$Mx = 0x = 0$$

A singular matrix has linearly dependent columns.

4.3.3 Symmetric matrices

Eigenvectors corresponding to distinct eigenvalues are orthogonal:

$$\begin{aligned}
\lambda\langle x, z\rangle = \langle Ax, z\rangle = \\
= (Ax)^T z = \qquad\qquad = x^T A^T z = \\
= x^T A z = \\
\langle x, Az\rangle = \\
= \mu\langle x, z\rangle
\end{aligned}$$

4.3.4 Eigen-decomposition

4.3.4.1 Raleigh quotient

$$Ax = \lambda x$$
$$\frac{x^T Ax}{x^T x} = \lambda \frac{x^T x}{x^T x} = \lambda$$

4.3.4.2 Finding eigenvector

To find the eigenvector you need to maximize the eigenvalue:

$$x = \max_v \frac{v^T Av}{v^T v}$$

After that you need to normalize it so the solution is invariant to rescaling:

$$x \leftarrow \frac{x}{||x||}$$

4.3.4.3 Deflating matrix

$$\bar{A} = A - \lambda x x^T$$

The deflation turns x into a zero eigenvalue eigenvector:

$$\bar{A}x = Ax - \lambda x x^T x$$
$$Ax - \lambda x = 0$$

Other eigenvalues are unchanged as eigenvectors with distinct eigenvalues are orthogonal because the matrix is symmetric:

$$\bar{A}z = Az - \lambda x x^T z$$
$$\bar{A}z = Az$$

4.3.4.4 Iterating

The maximization procedure is repeated on the deflated matrix until solution is zero. Minimization is iterated to get eigenvectors with negative eigenvalues. Eigenvectors with zero eigenvalues are obtained extending the obtained set to an orthonormal basis.

4.3.4.5 Eigen-decomposition

Let $V = [v_1 \dots v_n]$ be a matrix with orthonormal eigenvectors as columns. Let A be the diagonal matrix of corresponding eigenvalues. A square symmetric matrix can be diagonalizes as:

$$V^T AV = \Lambda$$

A diagonalized matrix is simpler to manage and has the same properties as the original one.

4.3.4.5.1 Proof

$$A[v_1 \dots v_n] = [v_1 \dots v_n \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{bmatrix}$$

$$AV = V\Lambda$$
$$V^{-1}AV = V^{-1}V\Lambda$$
$$V^TAV = \Lambda$$

V is a unitary matrix with orthonormal columns for which $V^{-1} = V^T$.

4.3.4.6 Positive semi-definite matrix

An $n \times n$ symmetric matrix M is positive semi-definite if all its eigenvalues are non-negative. Positive semi-definite:

- $\Leftrightarrow \forall x \in \mathbb{R}^n : x^TMx \geq 0$
- $\Leftrightarrow \exists B : M = B^TB$

4.3.4.7 Scaling transformation in standard basis

Let $x_i = [1, 0]$ and $x_2 = [0, 1]$ the standard orthonormal basis in $\mathbb{R}^2$. Let $x = [x_1, x_2]$ am arbitrary vector in $\mathbb{R}^2$. A linear transformation is a scaling transformation if it only stretches x along its directions.

4.3.4.8 Scaling transformation in eigenbasis

Let A be a non-scaling transformation in $\mathbb{R}$. Let $\{v_1, v_2\}$ be an eigenbasis for A. By representing vectors in $\mathbb{R}^2$ in terms of the $\{v_1, v_2\}$ basis A becomes a scaling transformation.

4.4 Principal component analysis

Principal component analysis or PCA is a non-supervised machine learning technique that accomplishes dimensionality reduction. Let X be a data matrix with correlated coordinates. PCA is a linear transformation mapping data to a system of uncorrelated coordinates. It corresponds to fitting an ellipsoid to the data, whose axis are the coordinates of the new space.

4.4.1 Procedure

Given a dataset $X \in \mathbb{R}^{n \times d}$ in d dimension: Compute the mean of the data:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Center the data into the origin:

$$X - \begin{bmatrix} \bar{x} \\ \dots \\ \bar{x} \end{bmatrix}$$

Compute the data covariance:

$$C = \frac{1}{n} X^T X$$

Compute the orthonormal eigen-decomposition of C:

$$V^T C V = A$$

Us it as the new coordinate system:

$$x' = V^{-1} x = V^T x$$

This method assumes linear correlation and Gaussian distribution.

4.4.2 Dimensionality reduction

Each eigenvalue corresponds to the amount of variance in that direction. Select only the k eigenvalues with largest eigenvalue for dimensionality reduction:

$$W = [v_1, \ldots, v_k]$$

$$x' = W^T x$$

Chapter 5

Probability theory

5.1 Discrete random variables

5.1.1 Probability mass function

Given a discrete random variable X taking values in $\mathcal{X} = \{v_1, \ldots, v_m\}$ its probability mass function $P : \mathcal{X} \to [0, 1]$ is defined as:

$$P(v_i) = Pr[X = v_i]$$

This function satisfies:

- $P(x) \geq 0$
- $\sum_{x \in \mathcal{X}} = 1$

5.1.2 Expected value

The expected value, mean or average of a random variable x is:

$$\mathbb{E}[x] = \mu = \sum_{x \in \mathcal{X}} xP(x) = \sum_{i=1}^{m} v_i P(v_i)$$

The expectation operator is linear:

$$\mathbb{E}[\lambda x + \lambda' y] = \lambda \mathbb{E}[x] + \lambda' \mathbb{E}[y]$$

5.1.3 Variance

The variance of a random variable is the moment of inertia of its probability mass function:

$$Var[x] = \sigma^2 = \mathbb{E}[(x - \mu)^2] = \sum_{x \in \mathcal{X}} (x - \mu)^2 P(x)$$

The standard deviation σ indicates the typical amount of deviation from the mean one should expect for a randomly drawn value for x.

5.1.4 Properties of mean and variance

5.1.4.1 Second moment

$$\mathbb{E}[x^2] = \sum_{x \in \mathcal{X}} x^2 P(x)$$

5.1.4.2 Variance in term of expectation

$$Var[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2$$

5.1.4.3 Variance and scalar multiplication

$$Var[\lambda x] = \lambda^2 Var[x]$$

5.1.4.4 Variance of uncorrelated variables

$$Var[x + y] = Var[x] + Var[y]$$

5.1.5 Probability distributions

5.1.5.1 Bernoulli distribution

The Bernoulli distribution indicates a variable for which there are two values: 1 for success and 0 for failure. Its parameter is p the probability of success. Its probability mass function:

$$P(x; p) = \begin{cases} p & if x = 1 \\ 1 - p & if x = 1 \end{cases}$$

- $\mathbb{E}[x] = p$
- $Var[x] = p(1 - p)$

5.1.5.1.1 Proof of mean

$$\begin{aligned} \mathbb{E}[x] &= \sum_{x \in \mathcal{X}} xP(x) \\ &= \sum_{x \in \{0;1\}} xP(x) \\ &= 0 \cdot (1 - p) + 1 \cdot p = p \end{aligned}$$

5.1.5.1.2 Proof of variance

$$\begin{aligned} Var[x] &= \sum_{x \in \mathcal{X}} (x - \mu)^2 P(x) \\ &= \sum_{x \in \{0;1\}} (x - \mu)^2 P(x) \\ &= (0 - p)^2 (1 - p) + (1 - p)^2 p \\ &= p^2 (1 - p) + (1 - p)(1 - p)p \\ &= (1 - p)(p^2 + p - p^2) \\ &= (1 - p)p \end{aligned}$$

5.1.5.2 Binomial distribution

The binomial distribution is the probability of a certain number of successes in n independent Bernoulli trials. Its parameters are p the probability of success and n the number of trials. Its probability mass function:

$$P(x; p, n) = \binom{n}{x} p^x (1 - p)^{n - x}$$

- $\mathbb{E}[x] = np$
- $Var[x] = np(1-p)$

5.1.6 Pairs of discrete random variables

5.1.6.1 Probability mass function

Given a pair of discrete random variables X and Y taking values $\mathcal{X} = \{v_1, \ldots, v_m\}$ and $\mathcal{Y} = \{w_1, \ldots, w_n\}$ the joint probability mass function is defined as:

$$P(v_i, w_j) = Pr[X = v_i, Y = w_j]$$

This satisfies:

- $P(x, y) \geq 0$
- $\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P(x, y) = 1$

5.1.6.2 Properties

5.1.6.2.1 Expected value

$$\mu_x = \mathbb{E}[x] = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} xP(x, y)$$

$$\mu_y = \mathbb{E}[y] = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} yP(x, y)$$

5.1.6.2.2 Variance

$$\sigma_x^2 = Var[(x - \mu_x)^2] = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} (x - \mu_x)^2 P(x, y)$$

$$\sigma_y^2 = Var[(y - \mu_y)^2] = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} (y - \mu_y)^2 P(x, y)$$

5.1.6.2.3 Covariance

$$\sigma_{xy} + \mathbb{E}[(x - \mu_x)(y - \mu_y)] = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} (x - \mu_x)(y - \mu_y) P(x, y)$$

5.1.6.2.4 Correlation coefficient

$$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

5.1.6.3 Multinomial distribution

Given n samples of an event with m possible outcomes, the multinomial distribution models the probability of a certain distribution of outcomes. It has parameters $p_1, \ldots, p_m$ probabiliry of each outcome and n the number of samples. Its probability mass function assumes $\sum_{i=1}^{m} x_i = n$ and it is:

$$P(x_1, \ldots, x_m; p_1, \ldots, p_m, n) = \frac{n!}{\prod_{i=1}^{m} x_i!} \prod_{i=1}^{m} p_i^{x_i}$$

- $\mathbb{E}[x_i] = np)i$
- $Var[x_i] = np_i(1 - p_i)$
- $Cov[x_i, x_j] = -np_ip_j$

:

$$P(x_1, \ldots, x_m; p_1, \ldots, p_m, n) = \frac{n!}{\prod_{i=1}^{m} x_i!} \prod_{i=1}^{m} p_i^{x_i}$$

- $\mathbb{E}[x_i] = np)i$
- $Var[x_i] = np_i(1 - p_i)$
- $Cov[x_i, x_j] = -np_ip_j$

5.2 Conditionally probability

The conditional probability is the probability of x once y is observed:

$$P(x|y) = \frac{P(x, y)}{P(y)}$$

Variables X and Y are statistical independent if and only if:

$$P(x, y) = P(x)P(y)$$

This implies:

- $P(x|y) = P(x)$
- $P(y|x) = P(y)$

5.2.1 Basic rules

5.2.1.1 Law of total probability

The marginal distribution of a variable is obtained from a joint distribution summing over all possible values of the other variable:

- $P(x) = \sum_{y \in \mathcal{Y}} P(x, y)$
- $P(y) = \sum_{x \in \mathcal{X}} P(x, y)$

5.2.1.2 Product rule

Conditional probabiliry definition implies that:

$$P(x, y) = P(x|y)P(y) = P(y|x)P(x)$$

5.2.2 Bayes' rule

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)}$$

This allows to invert statistical connection between effects x and cause y:

$$posterior = \frac{likelihood \times prior}{evidence}$$

The evidence can be obtained using the sum rule from likelihood and prior:

$$P(x) = \sum_y P(x, y) = \sum_y P(x|y)P(y)$$

5.3 Continuous random variables

5.3.1 Cumulative distribution function

To generalize the probability mass function to continuous domains it is considered the probability of intervals:

$$W = (a < X \leq b) \qquad A = (X \leq a) \qquad B = (X \leq b)$$

W and A are mutually exclusive, so:

$$P(B) = P(A) + P(W) \qquad P(W) = P(B) - P(A)$$

$F(q) = P(X \leq q)$ is the cumulative distribution function of X a monotonic function such that the probability of an interval is the difference:

$$P(a < X \leq b) = F(b) - F(a)$$

5.3.2 Probability density function

The derivative of the cumulative distribution function:

$$p(x) = \frac{d}{dx}F(x) \qquad F(q) = P(X \leq q) = \int_{-\infty}^{q} p(x)dx$$

So that it respect the properties:

- $p(x) \geq 0$
- $\int_{-\infty}^{\infty} p(x)dx = 1$

The probability density function of a value x can be greater than one provided the integral is one.

- $\mathbb{E}[x] = \mu = \int_{-\infty}^{\infty} xp(x)dx$
- $Var[x] = \sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 p(x)dx$

Definitions and formulae for discrete random variables carry over to continuous random variables with sums replaced by integrals.

5.3.3 Probability distribution

5.3.3.1 Gaussian or normal distribution

The normal distribution is a bell-shaped curved with parameters μ mean and σ^2 variance. Its probability density function is:

$$p(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Where:

- $\mathbb{E}[x] = \mu$
- $Var[x] = \sigma^2$

The standard normal distribution is $N(0,1)$. Every normal distribution can be transformed in a standard one:

$$z = \frac{x - \mu}{\sigma}$$

5.3.3.2 Beta distribution

The beta distribution is defined in the interval $[0,1]$ with parameters α and β. Its probability density function is:

$$p(x;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

Where:

- $\mathbb{E}[x] = \frac{\alpha}{\alpha+\beta}$
- $Var[x] = \frac{\alpha\beta}{(\alpha+\beta+1)}$
- $\Gamma(x+1) = x\Gamma(x)$
- $\Gamma(1) = 1$

It models the posterior distribution of parameter p of a binomial distribution after observing $n-1$ independent events with probability p and $\beta - 1$ with probability $1-p$.

5.3.3.3 Multivariate normal distribution

The multivariate normal distribution is the normal distribution for d-dimensional vectorial data. Its parameter are $\vec{\mu}$ mean vector and Σ covariance matrix. Its probability density function:

$$p(\vec{x},\vec{\mu},\Sigma) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(\vec{x}-\vec{\mu})^T\Sigma^{-1}(\vec{x}-\vec{\mu})}$$

Where:

- $\mathbb{E}[\vec{x}] = \vec{\mu}$
- $Var[\vec{x}] = \Sigma$

The standard measure of instance to mean is the squared Mahalanobis distance"

$$r^2 = (\vec{x}-\vec{\mu})^T\Sigma^{-1}(\vec{x}-\vec{\mu})$$

5.3.3.4 Dirichlet distribution

The Dirichlet distribution is defines in $x \in [0,1]^m$, $\sum_{i=1}^{m} x_i = 1$ It has parameters $\vec{\alpha} = \alpha_1,\ldots,\alpha_m$. Its probability density function is:

$$p(x_1,\ldots,x_m;\vec{\alpha}) = \frac{\Gamma((\alpha_0)}{\prod_{i=1}^{m}\Gamma(\alpha_i)} \prod_{i=1}^{m} x_i^{a_i-1}$$

Where $\alpha_0 = \sum_{j=1}^{m} \alpha_j$. Also:

- $\mathbb{E}[x_i] = \frac{\alpha_i}{\alpha_0}$
- $Var[x_i] = \frac{\alpha_i(\alpha_0 - \alpha_i)}{\alpha_0^2(\alpha_0+1)}$
- $Cps[x_i, x_j] = \frac{-\alpha_i \alpha_j}{a_0^2(\alpha_0+1)}$

This distribution models the posterior distribution of parameters p of a multinomial distribution after observing $\alpha_i - 1$ times each mutually exclusive event.

5.4 Probability laws

5.4.1 Expectation of an average

Consider a sample of $X_1, \ldots, X_n$ instances drawn from a distribution with mean μ and variance σ^2. Consider the random variable $\hat{X}_n$ measuring the sample average:

$$\hat{X}_n = \frac{X_1 + \cdots + X_n}{n}$$

Considering that $\mathbb{E}[a(X+Y)] = a(\mathbb{E}[x] + \mathbb{E}[y])$:

$$\mathbb{E}[\hat{X}_n] = \frac{1}{n}(\mathbb{E}[X_1] + \cdots + \mathbb{E}[X_n]) = p$$

5.4.2 Variance of an average

Consider the random variable $\hat{X}_n$ measuring the sample average. Its variance is computed as $Var[a(X+Y)] = a^2(Var[X] + Var[Y])$ if X and Y are independent:

$$Var[\hat{X}_n] = \frac{1}{n^2}(Var[X_1] + \cdots Var[X_n]) = \frac{\sigma^2}{n}$$

5.4.3 Chebyshev's inequality

Consider a random variable X with mean μ and variance σ^2. For all $a > 0$:

$$Pr[|X - \mu| \geq a] \leq \frac{\sigma^2}{a^2}$$

Considering $a = k\sigma$, for $k > 0$:

$$Pr[|X - \mu| \geq k\sigma] \leq \frac{1}{k^2}$$

Most of the probability mass of a random variable stays within few standard deviations from its mean.

5.4.4 Law of large numbers

Consider a sample $X_1, \ldots, X_n$ instances drawn from a distribution with mean μ and variance σ^2. For any $\varepsilon > 0$ its sample average $\hat{X}_n$ obeys:

$$\lim_{n \to \infty} Pr[|\hat{X}_n - \mu| > \varepsilon] = 0$$

It can be shown using Chebyshev's inequality and the facts that $\mathbb{E}[\hat{X}_n] = \mu$ and $Var[\hat{X}_n] = \frac{\sigma^2}{n}$:

$$Pr[|\hat{X}_n - \mathbb{E}[\bar{X}_n]| \geq \varepsilon] \leq \frac{\sigma^2}{n\varepsilon^2}$$

This tells that the accuracy of an empirical statistic increases with the number of samples.

5.4.5 Central limit theorem

Consider a sample of $X_1, \ldots, X_n$ instances drawn from a distribution with mean μ and variance σ^2. Regardless of the distribution of X, for $n \to \infty$ the distribution of the sample average $\hat{X}_n$ approaches a normal distribution. Its mean approaches μ and its variance $\frac{\sigma^2}{n}$. Thus the normalized sample average"

$$z = \frac{\hat{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Approaches a normal distribution $N(0, 1)$.

5.4.5.1 Interpretation

The sum of a sufficiently large sample of random measurements is approximately normally distributed. The form of their distribution can be arbitrary. This justifies the importance of the Normal distribution in real world applications.

5.5 Information theory

5.5.1 Entropy

Consider a discrete set of symbols $\mathcal{V} = \{v_1, \ldots, v_n\}$ with mutually exclusive probabilities. To design a binary code for each symbol minimizing the average length of messages, Shannon and Weaver proved that the optimal code assigns to each symbol a number of bits equal to $-\log P(v_i)$. The entropy of the set of symbols is the expected length of a message encoding a symbol assuming such optimal coding:

$$H[\mathcal{V}] = \mathbb{E}[-\log P(v)] = -\sum_{i=1}^{n} P(v_i) \log P(v_i)$$

5.5.2 Cross entropy

Consider two distributions P and Q over variable X. The cross entropy between P and Q measures the expected number of bits needed to code a symbol sampled from P using Q instead:

$$H(P, Q) = \mathbb{E}_P[-\log Q(v)] = -\sum_{i=1}^{n} P(v_i) \log Q(v_i)$$

It is often used as a loss for binary classification with P true distribution and Q the predicted one.

5.5.3 Relative entropy

Consider two distributions P and Q over variable X. The relative entropy or Kullback-Leibler divergence measures the expected length difference when coding instances sampled from P using Q instead.

$$\begin{aligned} D_{KL}(p||q) &= H(P,Q) - H(P) = \\ &= -\sum_{i=1}^{n} P(v_i) \log Q(v_i) + \sum_{i=1}^{n} P(v_i) \log P(v_i) = \\ &= \sum_{i=1}^{n} P(v_i) \log \frac{P(v_i)}{Q(v_i)} \end{aligned}$$

The KL-divergence is not a distance as it is not symmetric.

5.5.4 Conditional entropy

Consider two variables V and W with possibly different distributions P. The conditional entropy is the entropy remaining for variable W once V is known:

$$\begin{aligned} H(W|V) &= \sum_{v} P(v) H(W|V=v) = \\ &= -\sum_{v} P(v) \sum_{w} P(w|v) \log P(w|v) \end{aligned}$$

5.5.5 Mutual information

Consider two variables V and W with possibly different distributions P. The mutual information or information gain is the reduction in entropy for W once V is known:

$$\begin{aligned} I(W,V) &= H(W) - H(W|V) \\ &= -\sum_{w} p(w) \log p(w) + \sum_{v} P(v) \sum_{w} P(w|v) \log P(w|v) \end{aligned}$$

It is used in selecting the best attribute to use in building a decision tree, where V is the attribute and W is the label.

Chapter 6

Evaluation

6.1 Introduction

Evaluation requires to define the performance measures to be optimized. Performance of learning algorithms cannot be evaluated on the entire domain because of the generalization error, so there is a need for approximation. Performance evaluation is needed for:

- Tuning the hyperparameters of the learning method.
- Evaluating the quality of the learned predictor.
- Computing statistical significance of difference between different learning algorithms.

6.2 Performance measures

6.2.1 Training loss and performance measures

The training loss function measures the cost paid for predicting $f(x)$ for output y. It is designed to boost effectiveness and efficiency of learning algorithm. It is not necessarily the best measure of final performance, for example misclassification cost is never used as it is a piecewise constant no amenable to gradient descent. Multiple performance measures could be used to evaluate different aspects of a learner.

6.2.2 Binary classification

For binary classification the confusion matrix reports the effective label on the rows and the predicted one on the columns. Each entry contains the number of examples having label in row and predicted as column.

- TP true positives: positives predicted as positive.
- FP false positives: negative predicted as positive.
- TN true negative: negative predicted as negative.
- FN false negatives: positives predicted as negatives.

6.2.2.1 Accuracy

Accuracy is the fraction of correctly labelled examples among all predictions. It is one minus the misclassification cost:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}$$

For strongly unbalanced datasets it is not informative because predictions are dominated by the larger class and predicting everything as negative maximizes accuracy. One way of resolving this problem consists of introducing the rebalancing cost: a single positive counts as $\frac{N}{P}$ where $N = TN + FP$ and $P = TP + FN$.

6.2.2.2 Precision

Precision is the fraction of positives among examples predicted as positives. It measures the precision of the learner when predicting positive:

$$Pre = \frac{TP}{TP + FP}$$

6.2.2.3 Recall or sensitivity

Recall is the fraction of positive examples predicted as positives. It measures the coverage of the learner in returning positive examples:

$$Rec = \frac{TP}{TP + FN}$$

6.2.2.4 F-measure

Precision and recall are complementary: increasing prediction typically reduces recall. F-measures combines the two measures balancing the two aspects with a parameter β that defines the trade-off between precision and recall:

$$F_\beta = \frac{(1 + \beta^2)(Pre + Rec)}{\beta^2 Pre + Rec}$$

If $\beta = 1$ the measure is called F_1 and it is the harmonic mean of precision and recall:

$$F_1 = \frac{2(Pre + Rec)}{Pre + Rec}$$

6.2.2.5 Precision-recall curve

Classifiers often provide a confidence in the prediction and a hard decision is made setting a threshold on the classifier. Accuracy, precision, recall and F_1 all measures performance of a classifier for a specific threshold. It is possible to change the threshold if the interest is in maximizing a specific performance. By varying the threshold from minimum to maximum possible values we obtain a curve of performance measures. This curve can be shown plotting one measure like recall against the complementary one like precision. It is possible to investigate performance of the learner in different scenarios. The area under this curve can be used as a single aggregate value that combines performance of the algorithm for all possible threshold.

6.2.3 Multiclass classification

For multiclass classification the confusion matrix is a generalized evasion of the binary one such that n_{ij} is the number of examples with class y_i predicted as y_j. The main diagonal contains true positives for each class. The sum of off-diagonal elements along a column is the number of false positive for the column label and the sum of off-diagonal elements along a row is the number of false negatives for the raw label:

- $FP_i = \sum_{j \neq i} n_{ji}$
- $FN_i = \sum_{j \neq i} n_{ij}$

6.2.3.1 Muliclass accuracy

Accuracy, precision, recall and F_1 carry over to a per-class measure considering as negative examples from other classes:

- $Pre_i = \frac{n_{ii}}{n_{ii}+FP_i}$
- $Rec_i = \frac{n_{ii}}{n_{ii}+FN_i}$

Multiclass accuracy is the overall fraction of correctly classified examples:

$$MAcc = \frac{\sum_i n_{ii}}{\sum_i \sum_j n_i}$$

6.2.4 Regression

6.2.4.1 Root mean squared error

The root mean squared error for dataset $\mathcal{D}$ with $n = |\mathcal{D}|$ is:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(f(x_i) - y_i)^2}$$

6.2.4.2 Pearson correlation coefficient

The Pearson correlation coefficient for random variable X and Y is:

$$\rho = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{\mathbb{E}[(X - \bar{X})(Y - \bar{Y}]}{\sqrt{\mathbb{E}[(X - \bar{X})^2]\mathbb{E}[(Y - \bar{Y})^2]}}$$

For regression on $\mathcal{D}$ it becomes:

$$\rho = frac\sum_{i=1}^{n}(f(x_i) - \bar{f}(x_i))(y_i - \bar{y}_i)\sqrt{\sum_{i=1}^{n}(f(x_i) - \bar{f}(x_i))^2 \sum_{i=1}^{n}(y_i - \bar{y}_i)^2}$$

Where $\bar{z}$ is the average of z on $\mathcal{D}$.

6.2.4.3 Hold-out procedure

Computing the performance measure on training set would be optimistically biased, so there is a need to retain an independent set on which to compute performance:

- Validation set: when used to estimate performance of different algorithmic settings.
- Test set: when used to estimate final performance of selected model.

Hold-out procedure depends on the specific test and validation set chosen.

6.2.4.4 K-fold cross validation

In the k-fold cross validation $\mathcal{D}$ is split in k equal sized disjoint subsets $\mathcal{D}_i$. Then for $i \in [1, k]$L

1. Train predictor on $\mathcal{T} = \mathcal{D} \backslash \mathcal{D}_i$.
2. Compute score S of predictor $L(\mathcal{T}_i)$ on test set $\mathcal{D}_i$: $S_i = S_{\mathcal{D}_i}[L(\mathcal{T}_i)]$.

Then the average score across folds is returned:

$$S = \frac{1}{k} \sum_{i=1}^{k} S_i$$

6.2.4.4.1 Variance The variance of the average score is computed as:

$$Var[\bar{S}] = Var\left[\frac{S_1 + \cdots + S_k}{k}\right] = \frac{1}{k^2} \sum_{j=1}^{k} Var[S_j]$$

$Var[S_j]$ cannot be exactly computed, so it is approximated with the unbiased variance across folds"

$$Var[S_j] = Var[S_h] \approx \frac{1}{k-1} \sum_{i=1}^{k} (S_i - \bar{S})^2$$

Giving:

$$Var[\bar{S}] \approx \frac{1}{k^2} \sum_{i=1}^{k} (S_i - \bar{S})^2 = \frac{1}{k(k-1)} \sum_{i=1}^{k} (S_i - \bar{S})^2$$

6.3 Hypothesis testing

There is a need to compare generalization performance of two learning algorithms. There is a need to know whether observed difference in performance is statistically significant. Hypothesis testing allows to test the statistical significance of an hypothesis.

6.3.1 Test statistic

Given a null hypothesis H_0, the default hypothesis, for rejecting which evidence should be provided, a sample of k realizations of random variables $X_1, \ldots, X_k$, a test statistic is a statistic $T = h(X_1, \ldots, X_n$ whose value is used to decide whether to reject H_0 or not.

6.3.2 Glossary

- **Tail probability**: probability that T is at least as great (right tail) or at least as small (left tail) as the observed value t.
- **p-value**: probability of obtaining a value T at least as extreme as the observed t in case H_0 is true.
- **Type I error**: reject the null hypothesis when it's true.
- **Type II error**: accept the null hypothesis when it's false.
- **Significance level**: larges acceptable probability for committing a type I error.
- **Critical region**: set of values of T for which the null hypothesis is rejected.
- **Critical values**: values on the boundary of the critical region.

6.3.3 T-test

In the t-test the test statistics is given by the standardized or studentized mean:

$$T = \frac{\bar{X} - \mu_0}{\sqrt{\tilde{Var}[\bar{X}]}}$$

Where $\tilde{Var}[\bar{X}]$ is the approximated variance using the unbiased sample one. Assuming the samples come from an unknown normal distribution, the test statistics has a t_{k-1} distribution under the null hypothesis. The null hypothesis can be rejected at significance level α if:

$$T \leq -t_{k-1,\frac{\alpha}{2}} \qquad \vee \qquad T \geq t_{k-1,\frac{\alpha}{2}}$$

6.3.3.1 t_{k-1} distribution

The t_{k-1} distribution is a bell-shaped distribution similar to the Normal one. It is wider and shorter, reflecting greater variance due to using $\tilde{Var}[\bar{X}]$ instead of the true unknown variance of the distribution. $k-1$ is the number of degree of freedom of the distribution and it's related to the number of independent events observed. t_{k-1} tends to the standardized normal z for $k \to \infty$.

6.3.4 Comparing learning algorithms

6.3.4.1 Hypothesis testing

After having run k-fold cross validation procedure for algorithms A and B, the mean performance difference for them is computed:

$$\bar{\delta} = \frac{1}{k}\sum_{i=1}^{k}\delta_i = \frac{1}{k}\sum_{i=1}^{k} S_{\mathcal{D}_i}[L_A(\mathcal{T}_i)] - S_{\mathcal{D}_i}[L_B(\mathcal{T}_i)]$$

The null hypothesis is that the mean difference is zero.

6.3.4.2 T-test

At significance level α:

$$\frac{\bar{\delta}}{\sqrt{\tilde{Var}[\bar{\delta}]}} \leq -t_{k-1,\frac{\alpha}{2}} \qquad \vee \qquad \frac{\bar{\delta}}{\sqrt{\tilde{var}[\bar{\delta}]}} \geq t_{k-1,\frac{\alpha}{2}}$$

Where:

$$\sqrt{\tilde{Var}[\bar{\delta}]} = \sqrt{\frac{1}{k(k-1)} \sum_{i=1}^{k} (\delta_i - \bar{\delta})^2}$$

6.3.4.3 Notes

- If the two hypothesis where evaluated over identical samples the paired test is used.
- If no prior knowledge can tell the direction of difference the two-tailed test is used, otherwise the one-tailed test.

Chapter 7

Bayesian decision theory

7.1 Introduction

Bayesian decision theory allows to take optimal decisions in a fully probabilistic setting. It assumes all relevant probabilities are known, allowing to provide upper bounds on achievable errors and to evaluate classifiers accordingly. Bayesian reasoning can be generalized to cases when the probabilistic structure is not entirely known.

7.1.1 Input-output pairs

In binary classification assume that examples $(x, y) \in \mathcal{X} \times \{-1, 1\}$ are drawn from a known distribution $p(x, y)$. The task is predicting the class y of examples given x. This can be written in probabilistic terms using Bayes rule:

$$P(y|x) = \frac{p(x|y)P(y)}{p(x)}$$

7.1.1.1 Output given input

Bayes rule allows to compute the posterior probability given likelihood, prior and evidence:

$$posterior = \frac{likelihood \times prior}{evidence}$$

- The posterior $P(y|x)$ is the probability that class is y given that x was observed.
- The likelihood $p(x|y)$ is the probability of observing x given that its class is y.
- The prior $P(y)$ is the prior probability of the class without any evidence.
- The evidence $p(x)$ is the probability of the observation and following the law of total probability can be computed as:

$$p(x) = \sum_{i=1}^{2} p(x|y)P(y)$$

7.1.2 Expected error

The probability of error given x is:

$$P(error|x) = \begin{cases} P(y_2|x) & if\ y_1\ is\ chosen \\ P(y_1|x) & if\ y_2\ is\ chosen \end{cases}$$

The average probability of error is:

$$P(error) = \int_{-\infty}^{\infty} P(error|x)p(x)dx$$

7.1.3 Bayes decision rule

7.1.3.1 Binary case

$$y_B = \arg \max_{y_i \in \{-1,1\}} P(y_i|x) = \arg \max_{y_i \in \{-1,1\}} p(x|y_i)P(y_i)$$

7.1.3.2 Multiclass case

$$y_B = \arg \max_{y_i \in \{1,\ldots,c\}} P(y_i|x) = \arg \max_{y_i \in \{1,\ldots,c\}} p(x|y_i)P(y_i)$$

7.1.3.3 Optimal rule

The probability of error given x is:

$$P(error|x) = 1 - P(y_B|x)$$

The Bayes decision rule minimizes the probability of error.

7.2 Representing classifiers

7.2.1 Discriminant functions

A classifier can be represented as a set of discriminant functions $g_i(x), i \in 1, \ldots, c$, giving:

$$y = \arg \max_{i \in 1,\ldots,c} g_i(x)$$

A discriminant function is not unique and the most convenient one can be chosen for computational or explanatory reasons:

- $g_i(x) = P(y_i|x) = \frac{p(x|y_i)P(y_i)}{p(x)}$.
- $g_i(x) = p(x|y_i)P(y_i)$.
- $g_i(x) = \ln p(x|y_i) + \ln P(y_i)$.

7.2.2 Decision regions

The feature space is divided into decision regions $\mathcal{R}_1, \ldots, \mathcal{R}_x$ such that:

$$x \in \mathcal{R}_i \qquad if\ g_i(x) > g_j(x) \forall j \neq i$$

This regions are separated by decision boundaries, regions in which ties occur among the largest discriminant functions.

7.3 Multivariate normal density

The multivariate normal density is the function

$$d(\vec{x}) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(\vec{x}-\vec{\mu})^t \Sigma^{-1}(\vec{x}-\vec{\mu})}$$

The covariance matrix Σ is always symmetric and positive semi-definite and strictly positive definite is the dimension of the feature space is d.

7.3.1 Hyperellipsoids

The loci of constant density are hyperellipsoids of constant Mahalanobis distance from $\vec{x}$ to $\vec{\mu}$. The principal axes of such hyperellipsoids are the eigenvectors of Σ, their lengths are given by the corresponding eigenvalues.

7.3.2 Discriminant functions for normal density

Let the discriminant function be:

$$\begin{aligned} g_i(\vec{x}) &= \ln p(\vec{x}|y_i) + \ln P(y_i) = \\ &= -\frac{1}{2}(\vec{x}-\vec{\mu}_i)^t \Sigma_i^{-1}(\vec{x}-\vec{\mu}_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma_i| + \ln P(y_i) \end{aligned}$$

Discarding those terms which are independent of i:

$$g_i(x) = -\frac{1}{2}(\vec{x}-\vec{\mu}_i)^t \Sigma_i^{-1}(\vec{x}-\vec{\mu}_i) - \frac{1}{2}\ln|\Sigma_i| + \ln P(y_i)$$

7.3.2.1 Case $\Sigma_i = \sigma^2 I$

In the case $\Sigma_i = \sigma^2 I$ all features are statistically independent and all of them have the same variance σ^2. The covariance determinant $|\Sigma_i| = \sigma^{2d}$ can be ignored as being independent of i. The covariance inverse is given by $\Sigma_i^{-1} = \frac{1}{\sigma^2} I$. The discriminant function then becomes:

$$g_i(\vec{x}) = -\frac{||\vec{x}-\vec{\mu}_i||^2}{2\sigma^2} + \ln P(y_i)$$

Then, expanding the quadratic form:

$$g_i(\vec{x}) = -\frac{1}{2\sigma^2}[\vec{x}^t\vec{x} - 2\vec{\mu}_i^t\vec{x} + \vec{\mu}_i^t\vec{\mu}_i] + \ln P(y_i)$$

Discarding all the terms independent of i the linear discriminant function is obtained:

$$g_i(x) = \underbrace{\frac{1}{\sigma^2}\vec{\mu}_i^t}_{\vec{w}_i^t}\vec{x} \underbrace{- \frac{1}{2\sigma^2}\vec{\mu}_i^t\vec{\mu}_i + \ln P(y_i)}_{w_{i0}}$$

7.3.2.1.1 Separating hyperplane Setting $g_i(\vec{x}) = g_j(\vec{x})$ the decision boundaries are pieces of hyperplanes.

$$\underbrace{(\vec{\mu}_i - \vec{\mu}_j)^t}_{\vec{w}^t}\Bigg(\vec{x} - \Bigg(\underbrace{\frac{1}{2}(\vec{\mu}_i + \vec{\mu}_j) - \frac{\sigma^2}{||\vec{\mu}_i - \vec{\mu}_j||^2}\ln\frac{P(y_i)}{P(y_j)}(\vec{\mu}_i - \vec{\mu}_j)}_{\vec{x}_0}\Bigg)\Bigg)$$

This hyperplane is orthogonal to vector $\vec{w}$, the line linking the means and passes through $\vec{x}_0$: if the prior probabilities of the classes are equal, $\vec{x}_o$ is halfway between the means, otherwise it sifts away from the more likely mean.

7.3.2.1.1.1 Derivation of the separating hyperplane

$$\begin{aligned}
g_i(\vec{x}) - g_j(\vec{x}) &= 0 \\
\frac{1}{\sigma^2}\vec{\mu}_i^t\vec{x} - \frac{1}{2\sigma^2}\vec{\mu}_i^t\vec{\mu}_i + \ln P(y_i) - \frac{1}{\sigma^2}\vec{\mu}_j^t\vec{x} + \frac{1}{2\sigma^2}\vec{\mu}_j^t\vec{\mu}^j - \ln P(y_j) &= 0 \\
(\vec{\mu}_i - \vec{\mu}_j)^t\vec{x} - \frac{1}{2}(\vec{\mu}_i^t\vec{\mu}_i - \vec{\mu}_j^t\vec{\mu}_j) + \sigma^2\ln\frac{P(y_i)}{P(y_j)} &= 0 \\
\vec{w}^t(\vec{x} - \vec{x}_o) &= 0 \\
\vec{w} &= \vec{\mu}_i - \vec{\mu}_j \\
(\vec{\mu}_i - \vec{\mu}_j)^t\vec{x}_0 &= \frac{1}{2}(\vec{\mu}_i^t\vec{\mu}_i - \vec{\mu}_j^t\vec{\mu}_j) - \sigma^2\ln\frac{P(y_i)}{P(y_j)} \\
\vec{\mu}_i^t\vec{\mu}_i - \vec{\mu}_j^t\vec{\mu}_j &= (\vec{\mu}_i - \vec{\mu}_j)^t(\vec{\mu}_i + \vec{\mu}_j) \\
\ln\frac{P(y_i)}{P(y_j)} = \frac{(\vec{\mu}_i - \vec{\mu}_j)^t(\vec{\mu}_i - \vec{\mu}_j)}{(\vec{\mu}_i - \vec{\mu}_j)^t(\vec{\mu}_i - \vec{\mu}_j)}\ln\frac{P(y_i)}{P(y_j)} &= (\vec{\mu}_i - \vec{\mu}_j)^t\frac{\vec{\mu}_i - \vec{\mu}_j}{||\vec{\mu}_i - \vec{\mu}_j||^2}\ln\frac{P(y_i)}{P(y_j)} \\
\vec{x}_0 &= \frac{1}{2}(\vec{\mu}_i + \vec{\mu}_j) - \sigma^2\frac{\vec{\mu}_i - \vec{\mu}_j}{||\vec{\mu}_i - \vec{\mu}_j||^2}\ln\frac{P(y_i)}{P(y_j)}
\end{aligned}$$

7.3.2.2 Case $\Sigma_i = \Sigma$

In the case where all classes have the same covariance matrix the discriminant function become:

$$g_i(\vec{x}) = -\frac{1}{2}(\vec{x} - \vec{\mu}_i)^t\Sigma^{-1}(\vec{x} - \vec{\mu}_i) + \ln P(y_i)$$

Expanding the quadratic form and discarding terms independent of i we obtain the linear discriminant function:

$$g_i(\vec{x}) = \underbrace{\vec{\mu}_i^t\Sigma^{-1}}_{\vec{w}_i^t}\vec{x} \underbrace{- \frac{1}{2}\vec{\mu}_i^t\Sigma^{-1}\vec{\mu}_i + \ln P(y_i)}_{w_{i0}}$$

The separating hyperplanes are not necessarily orthogonal to the line linking the means:

$$\underbrace{(\vec{\mu}_i - \vec{\mu}_j)^t\Sigma^{-1}}_{\vec{w}^t}\Bigg(\vec{x} - \underbrace{\frac{1}{2}(\vec{\mu}_i + \vec{\mu}_j) - \frac{\ln\frac{P(y_i)}{P(y_j)}}{(\vec{\mu}_i - \vec{\mu}_j)^t\Sigma^{-1}(\vec{\mu}_i - \vec{\mu}_j)}(\vec{\mu}_i - \vec{\mu}_j)}_{\vec{x}_0}\Bigg))$$

7.3.2.3 Case Σ_i arbitrary

In the case where Σ_i is arbitrary the discriminant functions are inherently quadratic:

$$g_i(\vec{x}) = \vec{x}^t \underbrace{\left(-\frac{1}{2}\Sigma_i^{-1}\right)}_{W_i} \vec{x} + \underbrace{\vec{\mu}_i^t \Sigma_i^{-1}}_{\vec{w}_i^t} \vec{x} \underbrace{-\frac{1}{2}\vec{\mu}_i^t \Sigma_i^{-1} \vec{\mu}_i - \frac{1}{2}\ln|\Sigma_i| + \ln P(y_i)}_{w_{i0}}$$

In the two category case the decision surfaces are hyperquadratics like hyperplanes, pairs of hyperplanes, hyperspheres or hyperellipsoids.

7.4 Arbitrary inputs and outputs

7.4.1 Setting

Examples are input-output pairs $(x, y) \in \mathcal{X} \times \mathcal{Y}$ generated with probability $p(x, y)$.

7.4.2 Risk

7.4.2.1 Conditional risk

The conditional risk of predicting y^* given $\vec{x}$ is:

$$R(y^*|\vec{x}) = \int_{\mathcal{Y}} l(y^*, y) P(y|x) dy$$

7.4.2.2 Overall risk

The overall risk of a decision rule g is:

$$R[f] = \int R(f(x)|x) p(x) dx = \int_{\mathcal{X}} \int_{\mathcal{Y}} l(f(x), y) p(y, x) dx dy$$

7.4.3 Bayes decision rule

Bayes decision rule states:

$$y^B = \arg\min_{y \in \mathcal{Y}} R(y|x)$$

7.5 Handling features

7.5.1 Handling missing features - marginalize over missing variables

To marginalize over missing variables one must assume that input $\vec{x}$ consists of an observed part $\vec{x}_o$ and a missing part $\vec{x}_m$. The posterior probability of y_i given the observation can be obtained from probabilities over entire inputs by marginalizing over the missing part:

$$P(y_i|\vec{x}_o) = \frac{p(y_i,\vec{x}_o)}{p(\vec{x}_o)} = \frac{\int p(y_i,\vec{x}_o,\vec{x}_m)d\vec{x_m}}{p(\vec{x}_o)} =$$
$$= \frac{\int P(y_i|\vec{x}_o,\vec{x}_m)p(\vec{x}_o,\vec{x}_m)d\vec{x}_m}{\int p(\vec{x}_o,\vec{x}_m)d\vec{x}_m} =$$
$$= \frac{\int P(y_i|vecx)p(\vec{x})d\vec{x}_m}{\int p(\vec{x})d\vec{x}_m}$$

7.5.2 Handling noisy features - marginalize over true variables

To marginalize over true variables on must assume that $\vec{x}$ consists of a clean part $\vec{x}_c$ and a noisy part $\vec{x}_n$. Also there is a need for a noise model for the probability of the noisy feature given its true version $p(\vec{x}_n|\vec{x}_t)$. The posterior probability of y_i given the observation can be obtained from probabilities over clean inputs by marginalizing over true variables via the noise model:

$$P(y_i|\vec{x}_x,\vec{x}_n) = \frac{p(y_i,\vec{x}_c,\vec{x}_n)}{p(\vec{x}_c,\vec{x}_n)} = \frac{\int p(y_i,\vec{x}_c,\vec{x}_n,\vec{x}_t)d\vec{x}_t}{\int p(\vec{x}_c,\vec{x}_n,\vec{x}_t)d\vec{x}_t)} =$$
$$= \frac{\int p(y_i|\vec{x}_c m\vec{x}_n,\vec{x}_t)p(\vec{x}_c,\vec{x}_n,\vec{x}_t)d\vec{x}_t}{\int p(\vec{x}_c,\vec{x}_n,\vec{x}_t)d\vec{x}_t} =$$
$$= \frac{\int p(y_i|\vec{x}_c,\vec{x}_t)p(\vec{x}_n|\vec{x}_c,\vec{x}_t)p(\vec{x}_c,\vec{x}_t)d\vec{x}_t}{\int p(\vec{x}_n|\vec{x}_c,\vec{x}_t)p(\vec{x}_c,\vec{x}_t)d\vec{x}_t}$$
$$= \frac{\int p(y_i|\vec{x})p(\vec{x}_n|\vec{x_t})p(\vec{x})d\vec{x}_t}{\int p(\vec{x}_n|\vec{x}_t)p(\vec{x})d\vec{x}_t}$$

Chapter 8

Parameter estimation

8.1 Introduction

8.1.1 Setting

Data is sampled from a probability distribution $p(x, y)$, whose form is known but its parameters are unknown. The training set $\mathcal{D} = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ of examples sampled independent and identically distributed according to $p(x, y)$.

8.1.2 Task

The task is to estimate the unknown parameters of p from training data $\mathcal{D}$. This is the same as Bayesian decision theory: there is a need to compute the posterior probability of classes given examples, except the parameters of the distributions are unknown and a training set is provided instead.

8.1.3 Multi class classification

The training set can be divided into $\mathcal{D}_1, \ldots, \mathcal{D}_c$ subsets, one for each class such that $\mathcal{D}_i = \{\vec{x}_1, \ldots, \vec{x}_n\}$ contains independent and identically distributed examples for target class y_i. For any new example $\vec{x}$ the posterior probability of the class given the example and the full training set $\mathcal{D}$ is computed:

$$P(y_i|\vec{x}, \mathcal{D}) = \frac{p(\vec{x}|y_i, \mathcal{D})p(y_i|\mathcal{D})}{p(\vec{x}|\mathcal{D}}$$

8.1.3.1 Simplifications

$\vec{x}$ can be assumed independent of $\mathcal{D}_j$ with $j \neq i$ given y_i and $\mathcal{D}_i$. Without additional knowledge $p(y_i|\mathcal{D})$ can be computed as the fraction of examples with that class in the dataset. The normalizing factor $p(\vec{x}|\mathcal{D})$ can be computed marginalizing $p(\vec{x}|y_i, \mathcal{D}_i)p(y_i|\mathcal{D})$ over possible classes. There is a need to estimate class-dependent parameters $\vec{\theta_i}$ for $p(\vec{x}|y_i, \mathcal{D}_i$.

8.2 Maximum likelihood

Maximum likelihood or maximum a posteriori estimation assumes parameters $\vec{\theta_i}$ have fixed but unknown values. These are computed as those maximizing the probability of the observed examples $\mathcal{D}_i$. Obtained values are used to compute the probability for new examples:

$$p(\vec{x}|y_i, \mathcal{D}_i) \approx p(\vec{x}|\vec{\theta_i})$$

It assumes a prior distribution for the parameters $p(\vec{\theta_i})$ is available. This maximizes the likelihood of the parameters with respect to the training samples and there is no assumption about prior distributions for parameters.

$$\vec{\theta_i^*} = \arg\max_{\vec{\theta_i}} = p(\vec{\theta_i}|\mathcal{D}_i, y_i) = \arg\max_{\vec{\theta_i}} p(\mathcal{D}_i, y_i|\vec{\theta_i})p(\vec{\theta_i}) = \arg\max_{\vec{\theta_i}} p(\mathcal{D}_i, y_i|\vec{\theta_i})$$

8.2.1 Setting

A training data $\mathcal{D} = \{\vec{x}_1, \ldots, \vec{x}_n\}$ of independent and identically distributed examples for the target class y is available. Assuming the parameter vector $\vec{\theta}$ as a fixed but unknown value, the value maximizing its likelihood with respect to the training data is estimated:

$$\vec{\theta^*} = \arg\max_{\vec{\theta}} p(\mathcal{D}|\vec{\theta}) = \arg\max_{\vec{\theta}} \prod_{j=1}^{n} p(\vec{x}_j|\vec{\theta})$$

The joint probability over $\mathcal{D}$ decomposes into a product as examples are independent and identically distributed.

8.2.2 Maximizing log-likelihood

It is usually simpler to maximize the logarithm of the likelihood because of its monotonic nature:

$$\vec{\theta^*} = \arg\max_{\vec{\theta}} \ln p(\mathcal{D}|\vec{\theta}) = \arg\max_{\vec{\theta}} \sum_{j=1}^{n} \ln p(\vec{x}_j|\vec{\theta})$$

The necessary conditions for the maximum can be obtained zeroing the gradient with respect to $\vec{\theta}$:

$$\nabla_{\vec{\theta}} \sum_{j=1}^{n} \ln p(\vec{x}_j|\vec{\theta}) = \vec{0}$$

Points zeroing the gradient can be local or global maxima depending on the form of the distribution.

8.2.3 Univariate Gaussian case

For the univariate Gaussian with unknown μ and σ^2 the log likelihood is:

$$\mathcal{L} = \sum_{j=1}^{n} -\frac{1}{\sigma^2}(x_j - \mu)^2 - \frac{1}{2}\ln 2\pi\sigma^2$$

8.2.3.1 Mean

The gradient with respect to μ is:

$$\frac{\partial \mathcal{L}}{\partial \mu} = 2\sum_{j=1}^{n} -\frac{1}{2\sigma^2}(x_j - \mu)(-1) = \sum_{j=1}^{n} \frac{1}{\sigma^2}(x_j - \mu)$$

Setting the gradient to zero gives mean:

$$\sum_{j=1}^{n} \frac{1}{\sigma^2}(x_j - \mu) = 0 = \sum_{j=1}^{n}(x_j - \mu)$$
$$\sum_{j=1}^{n} x_j = \sum_{j=1}^{n} \mu$$
$$\sum_{j=1}^{n} = n\mu$$
$$\mu = \frac{1}{n}\sum_{j=1}^{n} x_j$$

8.2.3.2 Variance

The gradient with respect to σ^2 is:

$$\frac{\partial \mathcal{L}}{\partial \sigma^2} = \sum_{j=1}^{n} -(x_j - \mu)^2 \frac{\partial}{\partial \sigma^2}\frac{1}{2\sigma^2} - \frac{1}{2}\frac{1}{2\pi\sigma^2}2\pi =$$
$$= \sum_{j-1}^{n} -(x_j - \mu)^2 \frac{1}{2}(-1)\frac{1}{\sigma^4} - \frac{1}{2\sigma^2}$$

Setting the gradient to zero gives variance:

$$\sum_{j=1}^{n} \frac{1}{2\sigma^2} = \sum_{j=1}^{n} \frac{(x_j - \mu)^2}{2\sigma^4}$$
$$\sum_{j=1}^{n} \sigma^2 = \sum_{j=1}^{n}(x_j - \mu)^2$$
$$\sigma^2 = \frac{1}{n}\sum_{j=1}^{n}(x_j - \mu)^2$$

8.2.4 Multivariate Gaussian case

For the multivariate Gaussian with unknown $\vec{\mu}$ and Σ the log-likelihood is:

$$\sum_{j=1}^{n} -\frac{1}{2}(\vec{x}_j - \vec{\mu})^t \Sigma^{-1}(\vec{x}_j - \vec{\mu}) - \frac{1}{2}\ln(2\pi)^d|\Sigma|$$

The maximum-likelihood estimates are:

$$\vec{\mu} = \frac{1}{n}\sum_{j=1}^{n}\vec{x}_j$$

$$\Sigma = \frac{1}{n}\sum_{j=1}^{n}(\vec{x}_j - \vec{\mu})(\vec{x}_j - \vec{\mu})^t$$

8.2.4.1 Proof for the mean

The gradient with respect to the mean is:

$$\nabla\vec{\mu}\sum_{j=1}^{n} -\frac{1}{2}(\vec{x}_j - \vec{\mu})^t\Sigma^{-1}(\vec{x}_j - \vec{\mu}) - \frac{1}{2}\ln(2\pi)^d|\Sigma| = \sum_{j=1}^{n}\Sigma^{-1}(\vec{x}_j - \vec{\mu})$$

Noting that $\frac{\partial}{\partial\vec{x}}\vec{x}^T A\vec{x} = A^T\vec{x} + A\vec{x} = 2A\vec{x}$ for symmetric A. Setting the gradient to zero:

$$\sum_{j=1}^{n}\Sigma^{-1}(\vec{x}_j - \vec{\mu}) = \vec{0}$$

$$\sum_{j=1}^{n}(\vec{x}_j - \vec{\mu}) = \Sigma\vec{0} = \vec{0}$$

$$\sum_{j=1}^{n}\vec{x}_j = \sum_{j=1}^{n}\vec{\mu} = n\vec{\mu}$$

$$\vec{\mu} = \frac{1}{n}\sum_{j=1}^{n}\vec{x}_j$$

8.2.4.2 Proof for the covariance

The gradient with respect to the covariance is:

$$\frac{\partial}{\partial\Sigma}\sum_{j=1}^{n} -\frac{1}{2}(\vec{x}_j - \vec{\mu})^t\Sigma^{-1}(\vec{x}_j - \vec{\mu}) - \frac{1}{2}\ln(2\pi)^d|\Sigma| =$$

$$-\frac{1}{2}\left(\sum_{j=1}^{n}\frac{\partial}{\partial\Sigma}(\vec{x}_j - \vec{\mu})^t\Sigma^{-1}(\vec{x}_j - \vec{\mu}) + \sum_{j=1}^{n}\frac{\partial}{\partial\Sigma}\ln(2\pi)^d|\Sigma|\right)$$

Now, expanding the first derivative:

$$\frac{\partial}{\partial\Sigma}(\vec{x}_j-\vec{\mu})^t\Sigma^{-1}(\vec{x}_j-\vec{\mu}) =$$
$$(\vec{x}_j-\vec{\mu})(\vec{x}_j-\vec{\mu})^t\frac{\partial}{\partial\Sigma}\Sigma^{-1} =$$
$$-(\vec{x}_j-\vec{\mu})(\vec{x}_j-\vec{\mu})^t\Sigma^{-2}$$

Using the matrix derivative rule: $\frac{\partial}{\partial B}tr(ABC) = CA$, where $A = (\vec{x}_j-\vec{\mu})^t$, $B = \Sigma^{-1}$ and $C = (\vec{x}_j-\vec{\mu})$ and $tr(ABC) = ABC$ as ABC is scalar. Now, expanding the second derivative:

$$\frac{\partial}{\partial\Sigma}\ln(2\pi)^d|\Sigma| = \frac{1}{(2\pi)^d}|\Sigma|^{-1}\frac{\partial}{\partial\Sigma}(2\pi)^d|\Sigma| =$$
$$\frac{1}{(2\pi)^d}|\Sigma|^{-1}(2\pi)^d\frac{\partial}{\partial\Sigma}|\Sigma| = |\Sigma|^{-1}|\Sigma|\Sigma^{-1} = \Sigma^{-1}$$

Using the matrix derivative rule: $\frac{\partial}{\partial A}|A| = |A|A^{-1}$. Combining those results and putting them equal to zero:

$$-\frac{1}{2}\left(\sum_{j=1}^{n}\overbrace{-(\vec{x}_j-\vec{\mu})(\vec{x}_j-\vec{\mu})^t\Sigma^{-2}}^{\frac{\partial}{\partial\Sigma}(\vec{x}_j-\vec{\mu})^t\Sigma^{-1}(\vec{x}_j-\vec{\mu})}+\sum_{j=1}^{n}\overbrace{\Sigma^{-1}}^{\frac{\partial}{\partial\Sigma}\ln(2\pi)^d|\Sigma|}\right)=0$$

$$\sum_{j=1}^{n}\Sigma^{-1} = \sum_{j=1}^{n}(\vec{x}_j-\vec{\mu})(\vec{x}_j-\vec{\mu})^t\Sigma^{-2}$$
$$\Sigma^2\sum_{j=1}^{n}\Sigma^{-1} = \Sigma^2\sum_{j=1}^{n}(\vec{x}_j-\vec{\mu})(\vec{x}_j-\vec{\mu})^t\Sigma^{-2}$$
$$\sum_{j=1}^{n}\Sigma = \sum_{j=1}^{n}(\vec{x}_j-\vec{\mu})(\vec{x}_j-\vec{\mu})^t$$
$$n\Sigma = \sum_{j=1}^{n}(\vec{x}_j-\vec{\mu})(\vec{x}_j-\vec{\mu})^t$$
$$\Sigma = \frac{1}{n}\sum_{j=1}^{n}(\vec{x}_j-\vec{\mu})(\vec{x}_j-\vec{\mu})^t$$

8.2.5 General Gaussian case

In the case of a general Gaussian the maximum likelihood estimates for Gaussian parameters are their empirical estimates over the samples: the mean is the sample mean and the covariance matrix is the mean of the sample covariances.

8.3 Bayesian estimation

Bayesian estimation assumes parameters $\vec{\theta}_i$ are random variables with some known prior distribution. Observing examples turns prior distribution over parameters into a posterior distribution. Predictions for new examples are obtained integrating over all possible values for the parameters:

$$p(\vec{x}|y_i, \mathcal{D}_i) = \int_{\vec{\theta}_i} p(\vec{x}, \vec{\theta}_i|y_i, \mathcal{D}_i)d\vec{\theta}_i$$

8.3.1 Setting

Bayesian estimation assumes parameters $\vec{\theta}_i$ are random variables with some known prior distribution. Predictions for new examples are obtained integrating over all possible values for the parameters:

$$p(\vec{x}|y_i, \mathcal{D}_i) = \int_{\vec{\theta}_i} p(\vec{x}, \vec{\theta}_i|y_i, \mathcal{D}_i)d\vec{\theta}_i$$

Because probability of $\vec{x}$ given each class y_i is independent of the other classes y_i:

$$p(\vec{x}|\mathcal{D}) = \int_{\vec{\theta}} p(\vec{x}, \vec{\theta}|mathcalD)d\vec{\theta} = \int p(\vec{x}|\vec{\theta})p(\vec{\theta}|\mathcal{D})d\vec{\theta}$$

$p(\vec{x}|\vec{\theta})$ can be easily computed because the form and the parameters of the distribution are known, so there is a need to estimate the parameter posterior density given the training set:

$$p(\vec{\theta}|\mathcal{D}) = \frac{p(\mathcal{D}|\vec{\theta})p(\vec{\theta})}{p(\mathcal{D})}$$

$P(\mathcal{D})$ is a constant independent of $\vec{\theta}$ so it will not influence the final Bayesian decision, if the final probability is needed it can be computed:

$$P(\mathcal{D}) = \int_{\vec{\theta}} p(\mathcal{D}|\vec{\theta})p(\vec{\theta})d\vec{\theta}$$

8.3.2 Univariate normal case - unknown μ, known σ^2

In this case the examples are drawn from $p(x|\mu) \sim N(\mu, \sigma^2)$. The Gaussian mean prior distribution is itself normal: $p(\mu) \sim N(\mu_0, \sigma_0^2)$. The Gaussian mean posterior given the dataset is computed as:

$$p(\mu|\mathcal{D}) = \frac{p(\mathcal{D}|\mu)p(\mu)}{p(\mathcal{D}} = \alpha \prod_{j=1}^{n} p(x_j|\mu)p(\mu)$$

Where $\alpha = \frac{1}{p(\mathcal{D})}$ is independent of μ.

8.3.2.1 A posteriori parameter density

$$p(\mu|\mathcal{D}) = \alpha \prod_{j=1}^{n} \overbrace{\frac{1}{\sqrt{2\pi}\sigma} e^{[-\frac{1}{2}(\frac{x_j-\mu}{\sigma})^2]}}^{p(x_j|\mu)} \overbrace{\frac{1}{\sqrt{2\pi}\sigma_0} e^{[-\frac{1}{2}(\frac{\mu-\mu_0}{\sigma_0})^2]}}^{p(\mu)} =$$

$$= \alpha' e^{[-\frac{1}{2}(\sum_{j=1}^{n}(\frac{\mu-x_j}{\sigma})^2+(\frac{\mu-\mu_0}{\sigma_0})^2)]} =$$

$$= \alpha'' e^{[-\frac{1}{2}[(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2})\mu^2-2(\frac{1}{\sigma^2}\sum_{j=1}^{n}x_j+\frac{\mu_0}{\sigma_0^2})\mu]]}$$

Where the normal distribution:

$$p(\mu|\mathcal{D}) = \frac{1}{\sqrt{2\pi}\sigma} e^{[-\frac{1}{2}(\frac{\mu-\mu_n}{\sigma^n})^2]}$$

8.3.2.2 Recovering mean and variance

$$\left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}\right)\mu^2 - 2\left(\frac{1}{\sigma^2}\sum_{j=1}^{n} x_j + \frac{\mu_0}{\sigma_0^2}\right)\mu + \alpha''' = \left(\frac{\mu-\mu_n}{\sigma_n}\right)^2$$
$$= \frac{1}{\sigma_n^2}\mu^2 - 2\frac{\mu_n}{\sigma_n^2}\mu + \frac{\mu_n^2}{\sigma_n^2}$$

Solving for μ_n and σ_n^2:

$$\mu_n = \left(\frac{n\sigma_0^2}{n\sigma_o^2 + \sigma^2}\right)\bar{\mu}_n + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2}\mu_0 \qquad \sigma_n^2 = \frac{\sigma_0^2\sigma^2}{n\sigma_0^2 + \sigma^2}$$

Where the sample mean $\bar{\mu}_n = \frac{1}{n}\sum_{j=1}^{n} x_j$.

8.3.2.3 Interpreting the posterior

The mean is a linear combination of the prior μ_0 and the sample means $\bar{\mu}_n$. The more training examples the more sample mean dominates over the prior mean. The more training examples, the more variance decreases making the distribution sharply peaked over its mean:

$$\lim_{n\to\infty} \frac{\sigma_0^2\sigma^2}{n\sigma_0^2 + \sigma^2} = \lim_{n\to\infty} \frac{\sigma^2}{n} = 0$$

8.3.2.4 Computing the class conditional density

$$p(x|\mathcal{D}) = \int p(x|\mu)p(\mu|\mathcal{D})d\mu =$$
$$= \int \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} \frac{1}{\sqrt{2\pi}\sigma_n} e^{-\frac{1}{2}(\frac{\mu-\mu_n}{\sigma_n})^2} d\mu =$$
$$\sim N(\mu_n, \sigma^2 + \sigma_n^2)$$

The probability of x given the dataset for the class is a Gaussian with mean equal to the posterior mean, variance equal to the sum of the known variance σ^2 and an additional variaci σ_n^2 due to the uncertainty of the mean.

8.3.3 Multivariate normal case - unknown μ, known Σ

This is a generalization of the univariate case:

- $p(\vec{x}|\vec{\mu}) \sim N(\vec{\mu}, \Sigma)$
- $p(\vec{\mu}) \sim N(\vec{\mu}_0, \Sigma_0)$
- $p(\vec{\mu}|\mathcal{D}) \sim N(\vec{\mu}_n, \Sigma_n)$
- $p(\vec{x}|\mathcal{D}) \sim N(\vec{\mu}_n, \Sigma + \Sigma_n)$

8.3.4 Gamma distribution

The gamma distribution is defined in the interval $[0, \infty]$. Its parameters are $\alpha > 0$ the shape and $\beta > 0$ the rate. Its probability density function is:

$$p(x; \alpha, \beta) = \frac{\mu^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$$

With $\mathbb{E}[x] = \frac{\alpha}{\beta}$ and $Var[x] = \frac{\alpha}{\beta^2}$. It is used to model the prior distribution of the precision.

8.3.5 Univariate normal case - unknown μ and $\lambda = \frac{1}{\sigma^2}$

In this case the samples are drawn from $p(x|\mu, \lambda) \sim N(\mu, \frac{1}{\lambda}$. The prior of mean and precision is the Normal Gamma distribution:

$$\begin{aligned} p(\mu, \lambda) &= p(\mu|\lambda)p(\lambda) = N\left(\mu|\mu_0, \frac{1}{k_0\lambda}\right) Ga(\lambda|\alpha_0, \beta_0) = \\ &= NG(\mu, \lambda|\mu_0, k_0, \alpha_0, \beta_0) \end{aligned}$$

8.3.5.1 A posteriori parameter density

$$p(\mu, \lambda|\mathcal{D}) = \frac{1}{\mathcal{D}} \prod_{j=1}^{n} \overbrace{\frac{\lambda^{\frac{1}{2}}}{\sqrt{2\pi}} e^{-\frac{\lambda}{2}(x_j-\mu)^2}}^{p(x_j|\mu,\lambda)} \overbrace{\frac{(k_0\lambda)^{\frac{1}{2}}}{\sqrt{2\pi}} e^{-\frac{k_0\lambda}{2}(\mu-\mu_0)^2}}^{p(\mu|\lambda)} \overbrace{\frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} \lambda^{\alpha_0-1} e^{-\beta_0\lambda}}^{p(\lambda)} \propto$$

$$\propto \lambda^{\alpha_0+\frac{n}{2}-1} e^{-\beta_0\lambda} \lambda^{\frac{1}{2}} e^{\frac{\lambda}{2}[\sum_{j=1}^{n}(x_j-\mu)^2 - k_0(\mu-\mu_0)^2}$$

The posteriori parameter density is still a Normal Gamma distribution:

$$p(\mu, \lambda|\mathcal{D}) = NG(\mu, \lambda|\mu_n, k-n, \alpha_n, \beta_n)$$

Where:

- $\mu_n = \frac{k_0\mu_0 + n\bar{\mu}_n}{k_0+n}$, the weighted average of prior μ_0 and sample means μ_n, weighted by k_0 and n respectively.
- $k_n = k_0 + n$, increased by the number of samples.
- $\alpha_n = \alpha_0 + \frac{n}{2}$ increased by half the number of samples.
- $\beta_n = \beta_0 + \frac{1}{2}\sum_{j=1}^{n}(x_j - \bar{\mu}_n)^2 + \frac{k_0 n(\bar{\mu}_n - \mu_0)^2}{2(k_0+n)}$, this is the sum of prior sum of sqares β_0 and sample um of squares and a term due to the discrepancy between the sample mean and the prior mean.

8.3.5.2 Computing the posterior predictive

$$p(x|\mathcal{D}) = \int_\mu \int_\lambda p(x|\mu,\lambda)p(\mu,\lambda|\mathcal{D})d\mu d\lambda$$
$$= \frac{P(x,\mathcal{D})}{P(\mathcal{D})} = t_{2\alpha_n}\left(x|\mu_n, \frac{\beta_n(k_n+1)}{\alpha_n k_n}\right)$$

This is a T-distribution with mean μ_n and precision $\frac{\beta_n(k_n+1)}{\alpha_n k_n}$.

8.3.6 Wishart distribution

The Wishart distribution is defined over $d \times d$ a positive semi-definite matrix. Its parameters are $\nu > d-1$, the degrees of freedon adn $T > 0$ the $d \times d$ scale matrix. Its probability density function is:

$$p(X;\nu,T) = \frac{1}{2^{\nu\frac{d}{2}}|T|^{\frac{\nu}{2}}\Gamma_d(\frac{\nu}{2})}|X|^{\frac{\nu-d-1}{2}}e^{-\frac{1}{2}tr(T^{-1}X)}$$

With $\mathbb{E}[X] = \nu T$ and $Var[X_{ij}] = \nu(T_i i^2 + T_{ii}T_{jj})$. It is used to model the prior distribution of the precision matrix. T is the prior covariance.

8.3.7 Multivariate normal case - unknown μ and Σ

In this case the examples are drawn from $p(\vec{x}|\vec{\mu},\Lambda) \sim N(\vec{\mu},\Lambda^{-1})$. The prior of mean and precision is the Normal Wishart distribution:

$$p(\vec{\mu},\Lambda) = p(\vec{\mu}|\Lambda)p(\Lambda) = N(\vec{\mu}|\vec{\mu}_0,(k_0\Lambda)^{-1})Wi(\Lambda|\nu,T)$$

8.3.7.1 A posteriori parameter density

$$p(\vec{\mu},\Lambda) = N(\vec{\mu}|\vec{\mu}_0,(k_0\Lambda)^{-1})Wi(\Lambda|\nu,T)$$

Where:

- $\mu_n = \frac{k_0\mu_0+n\bar{\mu}_n}{k_0+n}$
- $T_n = T + \sum_{i=1}^{n}(x_i-\bar{\mu}_n)(x_i-\bar{\mu}_n)^T + \frac{kn}{k+n}(\mu_0-\bar{\mu}_n)(\mu_0-\bar{\mu})n)^T$
- $\nu_n = \nu + n$
- $k_n = k + n$

8.3.7.2 Computing the posterior predictive

$$p(x|\mathcal{D}) = t_{\nu-d+1}\left(x|\mu_n, \frac{T_n(k_n+1)}{k_n(\nu_n-d+1)}\right)$$

8.4 Sufficient statistics

8.4.1 Definition

Any function on a set of samples $\mathcal{D}$ is a statistic. A statistic $s = \phi(\mathcal{D})$ is sufficient for some parameters $\vec{\theta}$ if $P(\mathcal{D}|\vec{s}, \vec{\theta}) = P(\mathcal{D}|\vec{s})$. If $\vec{\theta}$ is a random variable, a sufficient statistic contains all relevant information $\mathcal{D}$ has for estimating it:

$$p(\vec{\theta}|\mathcal{D}, \vec{s}) = \frac{p(\mathcal{D}|\vec{\theta}, \vec{s})p(\vec{\theta}|\vec{s})}{p(\mathcal{D}|\vec{s})} = p(\vec{\theta}|\vec{s})$$

A sufficient statistic allows to compress a sample $\mathcal{D}$ into fewer values. Sample mean and covariance are sufficient statistics for true mean and covariance of the Gaussian distribution.

8.4.2 Conjugate priors

Given a likelihood function $p(\vec{x}|\vec{\theta})$ and a prior distribution $p(\vec{\theta})$, $p(\vec{\theta})$ is a conjugate prior for $p(\vec{x}|\vec{\theta})$ if the posterior distribution $p(\vec{\theta}|\vec{x})$ is in the same family as the prior $p(\vec{\theta})$

8.5 Bernoulli distribution

8.5.1 Setting

The Bernoulli distribution represent a boolean event with $x = 1$ for success and $x = 0$ for failure. Its parameter is θ, the probability of success. Its probability mass function is:

$$P(x|\theta) = \theta^x(1-\theta)^{1-x}$$

Its beta conjugate prior is:

$$P(\theta|\psi) = P(\theta|\alpha_h, \alpha_t) = \frac{\Gamma(\alpha)}{\Gamma(\alpha_h)\Gamma(\alpha_t)}\theta^{\alpha_h-1}(1-\theta)^{\alpha_t-1}$$

8.5.2 Maximum likelihood estimation

Let $\mathcal{D} = \{H, H, T, T, T, H, H\}$ of N realizations. Its likelihood function is:

$$p(\mathcal{D}|\theta) = \theta \cdot \theta \cdot (1-\theta) \cdot (1-\theta) \cdot (1-\theta) \cdot \theta \cdot \theta = \theta^h(1-\theta)^t$$

Its maximum likelihood parameter:

$$\begin{aligned}
\frac{\partial}{\partial\theta}\ln p(\mathcal{D}|\theta) = 0 \Rightarrow \frac{\partial}{\partial\theta} h\ln\theta + t\ln(1-\theta) &= 0 \\
h\frac{1}{\theta} - t\frac{1}{1-\theta} &= 0 \\
h(1-\theta) &= t\theta \\
\theta &= \frac{h}{h+t}
\end{aligned}$$

h and t are the sufficient statistics.

8.5.3 Bayesian estimation

Parameter posterior is proportional to:

$$P(\theta|\mathcal{D},\psi) \propto P(\mathcal{D}|\theta)P(\theta|\psi) \propto \theta^t(1-\theta)^t\theta^{\alpha_h-1}(1-\theta)^{\alpha_t-1}$$

The posterior has a beta distribution with parameters $h+\alpha_h$ and $t+\alpha_t$:

$$P(\theta|\mathcal{D},\psi) \propto \theta^{h+\alpha_h-1}(1-\theta)^{t+\alpha_t-1}$$

The prediction for a new event is the expected value of the posterior beta:

$$\begin{aligned} P(x|\mathcal{D}) &= \int P(x|\theta)P(\theta|\mathcal{D},\psi)d\theta & &= \int \theta P(\theta|\mathcal{D},\psi)d\theta \\ &= \mathbb{E}_{P(\theta|\mathcal{D},\psi)}[\theta] & &= \frac{h+\alpha_h}{h+t+\alpha_h+\alpha_t} \end{aligned}$$

8.5.3.1 Interpreting priors

The prior knowledge is encoded as the number $\alpha = \alpha_h + \alpha_t$ of imaginary experiments. It is assumed α_t times heads was observed. α is the equivalent sample size and $\alpha \to 0$ reduces the estimation to the classical maximum likelihood approach.

8.6 Multinomial distribution

8.6.1 Setting

The multinomial distribution models categorical event with r states $x \in \{x^1,\dots,x^r\}$. One-hot encoding $\vec{z}(x) = [z_1(x),\dots,z_r(x)]$ with $z_k(x) = 1$ if $x = x^k$, o otherwise. Its parameters are $\vec{\theta} = [\theta_1,\dots,\theta_r]$ the probability of each state. Its probability mass function is:

$$P(x|\vec{\theta}) = \prod_{k=1}^{r} \theta_k^{z_k(x)}$$

Its Dirichlet conjugate prior:

$$P(\vec{\theta}|\psi) = P(\vec{\theta}|\alpha_1,\dots,\alpha_r) = \frac{\Gamma(\alpha)}{\prod_{k=1}^{r}\Gamma(\alpha_k)} \prod_{k=1}^{r} \theta_k^{\alpha_k-1}$$

8.6.2 Maximum likelihood estimation

Let $\mathcal{D}$ a dataset of N realizations. The likelihood function is:

$$p(\mathcal{D}|\vec{\theta} = \prod_{j=1}^{N}\prod_{k=1}^{r} \theta_k^{z_k(x_j)} = \prod_{k=1}^{r} \theta_k^{N_k}$$

The maximum likelihood parameter is $\theta_k = \frac{M_k}{N}$ and $N_1,\dots,N_r$ are the sufficient statistics.

8.6.3 Bayesian estimation

The parameter posterior is proportional to:

$$P(\vec{\theta}|\mathcal{D},\psi) \propto P(\mathcal{D}|\vec{\theta})P(\theta|\psi) \propto \prod_{k=1}^{r} \theta_k^{N_k}\theta_k^{\alpha_k-1}$$

The posterior has a Dirichlet distribution with parameters $N_k + \alpha_k$, where $k = 1, \ldots, r$"

$$P(\vec{\theta}|\mathcal{D},\psi) \propto \prod_{k=1}^{r} \theta_k^{N_k+\alpha_k-1}$$

The prediction for a new event is the expected value of the posterior Dirichlet:

$$P(x_k|\mathcal{D}) = \int \theta_k P(\vec{\theta}|\mathcal{D},\psi)d\vec{\theta} = \mathbb{E}_{P(\vec{\theta}|\mathcal{D},\psi)}[\theta_k] = \frac{N_k+\alpha_k}{N+\alpha}$$

Chapter 9

Bayesian networks

9.1 Inference in graphical models

Assume that there is evidence e on the state of a subset of variables W in the model. Inference amounts at computing the posterior probability of a subset X of the non-observed variables given the observations: $p(X|E = e)$.

9.1.1 Efficiency

The posterior probability can alwasy be computed as the ratio of two joint probabilities:

$$p(X|E = e) = \frac{p(X, E = e)}{p(E = e)}$$

The problem consists of estimating such joint porbabilities when dealing with a large number of variables. Directly working on the full joint probabilities requires time exponential in the number of variables: if all N variables are discrete and take one of K possible values, a joint probability table has K^N entries. There is a way to exploit the structure in graphical models to do inference more efficiently.

9.1.2 Inference on a chain

Let a probability chain be:

$$p(X) = p(X_1)p(X_2|X_1)p(X_3|X_2) \cdot p(X_N|X_{N-1})$$

The marginal probability of an arbitrary X_n is:

$$p(X_n) = \sum_{X_1} \cdots \sum_{X_{n-1}} \sum_{X_{n+1}} \cdot \sum_{X_N} p(X)$$

Only the $p(X_n|X_{N-1})$ is involved in the last summation which can be computed first, giving a function of X_{N-1}:

$$\mu_\beta(X_{N-1}) = \sum_{X_N} p(X_N|X_{N-1})$$

The marginalization can be iterated as:

$$\mu_\beta(X_{N-2}) = \sum_{X_{N-1}} p(X_{N-1}|X_{N-2})\mu_\beta(X_{N-1})$$

Down to the desired variable X_n, giving:

$$\mu_\beta(X_n) = \sum_{X_{n+1}} p(X_{n+1}|X_n)\mu_\beta(X_{n+1})$$

The same procedure can be applied starting from the other end of the chain, giving:

$$\mu_\alpha(X_2) = \sum_{X_1} p(X_1)p(X_2|x_1)$$

Up to $\mu_\alpha(X_n)$. The marginal probability is now computed as the product of the contributions coming from both ends:

$$p(X_n) = \mu_\alpha(X_n)\mu_\beta(X_n)$$

9.1.3 Inference as message passing

$\mu_\alpha(X_n)$ can be though as a message passing from X_{n-1} to X_n:

$$\mu_\alpha(X_n) = \sum_{X_{n-1}} p(X_n|X_{n-1})$$

$\mu_\beta(X_n)$ can be thought as a message passing from X_{n+1} to X_n:

$$\mu_\beta(X_n) = \sum_{X_{n+1}} p(X_{n+1}|X_n)\mu_\beta(X_{n+1})$$

Each outgoing message is obtained multiplying the incoming message by the local probability and summing over the node values.

9.1.4 Full message passing

Let's suppose we want to know marginal probabilities for a number of different variables X_i:

1. A message is sent from $\mu_\alpha(X_1)$ up to $\mu_\alpha(X_N)$.
2. A message is set from $\mu_\beta(X_N)$ down to $\mu_\beta(X_1)$.

If all nodes store messages any marginal probability can be computed as:

$$p(X_i) = \mu_\alpha(X_i)\mu_\beta(X_i)$$

For any i having sent a double number of messages with respect to a single marginal computation.

9.1.5 Adding evidence

If some nodes X_e are observed their observed values are used instead of summing over all possible values when computing their messages.

9.1.6 Computing conditional probability given evidence

When adding evidence, the message passing procedure computes the joint probability of the variable and the evidence, and it has to be normalized to obtain the conditional probability given the evidence:

$$p(X_n|X_e = x_e) = \frac{p(X_n, X_e = x_e)}{\sum\limits_{X_n} p(X_n, X_e = x_e)}$$

9.1.7 Inference on trees

Efficient inference can be computed for the broad family of tree-structured models: undirected trees, directed trees and directed polytrees.

9.2 Factor graphs

9.2.1 Description

Efficient inference algorithms can be better explained using an alternative graphical representation called factor graph. A factor graph is a graphical representation of a graphical model highlighting its factorizations (conditional probabilities). The factor graph has one node for each node in the original graph, and one additional of a different type for each factor. A factor node has undirected links to each of the node variables in the factor.

9.2.2 Sum-product algorithm

The sum-product algorithm is an efficient algorithm for exact inference on tree-structured graphs. It is a message passing algorithm as its simpler version for chain. It can be applied to undirected models and directed ones.

9.2.2.1 Computing marginals

The marginal probability of X is $p(X) = \sum\limits_{x \setminus X} p(X)$. Generalizing the message passing scheme, this can be computed as the product of messages coming from all neighbouring factors f_s:

$$p(X) = \prod_{f_s \in ne(X)} \mu_{f_s \to X}(X)$$

9.2.2.2 Factor messages

Each factor message is the product of messages coming from nodes other than X, times the factor, summed over all possible values of the factor variables other than X:

$$\mu_{f_s \to X}(X) = \sum_{X_1} \cdots \sum_{X_M} f_s(X, X_1, \ldots, X_M) \prod_{X_m \in ne(f_s) \setminus X} \mu_{X_m \to f_s}(X_m)$$

9.2.2.3 Node messages

Each message from node X_m to factor f_s is the product of the factor messages to X_m coming from factors other than f_s:

$$\mu_{X_m \to f_s}(X_m) = \prod_{f_l \in ne(X_m) \setminus f_s} \mu_{f_l \to X_m}(X_m)$$

9.2.2.4 Initialization

Message passing start from leaves, factors or nodes. Messages from leaf factors are initialized to the factor itself: $\mu_{f \to x}(x) = f(x)$. Messages from leaf nodes are initialized to 1: $\mu_{x \to f}(x) = 1$.

9.2.2.5 Message passing scheme

The node X whose marginal has to be computed is designed as root. Messages are sent from all leaves to their neighbours. Each internal node sends its message towards the root as soon as it received messages from all other neighbours. Once the root as collected all messages, the marginal can be computed as the product of them.

9.2.2.6 Full message passing scheme

In order to be able to compute marginals for any node, messages need to pass in all directions:

1. Choose an arbitrary node as root.
2. Collect messages for the root starting from leaves.
3. Send messages from the root down to the leaves.

All messages are passed in all directions using only twice the number of computations used for a single marginal.

9.2.2.7 Adding evidence

If some nodes X_e are observed, their values are used instead of summing over all possible values when computing their messages. After normalization, this gives the conditional probability given the evidence.