Committee machines and Principal Component Analysis (PCA)

Mikko Parviainen

November 21, 2012
Committee Machines

- Divide and Conquer principle: a complex computational task is divided into computationally simple tasks and combining the solutions of those tasks.
- A supervised learning scheme is employed in which individual experts try to learn the same task.
- The combination of experts constitutes the committee.
- It is likely that training time would be longer if the committee was replaced by a single (big) network with equal number of parameters.
Types of committee machines

- There are two basic types: static structures and dynamic structures.
- The responses of several structures are combined by means of a mechanism that does not involve input signal.
- In dynamic structures the input signal is directly involved in the process of combining the experts’ output.
Architecture of committee machines

\[ x(n) \rightarrow \text{Expert 1} \rightarrow y_1(n) \]
\[ \text{Expert 2} \rightarrow y_2(n) \]
\[ \text{Expert K} \rightarrow y_K(n) \]
\[ \text{Combiner} \rightarrow y(n) \]

Input: \( x(n) \)
Output: \( y(n) \)
Solid lines represent static structures.

If there is a connection from the input to the output, then the structure is dynamic.

In dynamic structures the values of input signals can be used e.g. gating the outputs of individual experts.
Static structures I

Ensemble averaging

- The outputs of experts are linearly combined.
- The belief: the combination of several neural networks is likely to take less time to train than a single large neural network.
- The individual networks have less parameters than a large neural network $\Rightarrow$ overfitting is more likely to be avoided.
- Overfitting occurs when the cardinality of parameters is comparable to the number of data samples in the training set.
Static structures II

Boosting

- In boosting scheme a weak learning algorithm is converted to possess an arbitrarily high accuracy.
- A weak learning algorithm is any learning algorithm that is better than a random one i.e. the prediction error probability is less than $1/2^*$ i.e. not so good.
- Contrast to Ensemble averaging, in boosting experts are trained with different data sets.
- The data sets may have completely different distributions.
- Boosting in general can be used to improve the performance of any learning algorithm.
Static structures III

- Boosting by filtering:
  Training samples are filtered with different instances of a weak learning algorithm. The term filtering comes from the fact that different subsets of training data are created; one expert filters its input to create another training data set for other experts.
Static structures IV

► Boosting by subsampling:
A training sample set size is fixed. The examples are "resampled" according to a given pdf during training.

► Boosting by reweighting:
A training sample set size is fixed. The approach assumes that the weak learning algorithm can receive "weighted" examples.

► Boosting by filtering and boosting by subsampling (AdaBoost) are covered in the course book.

*Pavel, Křížek, "Feature selection based on the training set manipulation", 2005
Principal Component Analysis (PCA) I

- In PCA the fundamental idea is data dimensionality reduction in such a way that the most of the intrinsic information is retained.
- More specifically, PCA ensures that the rate of decrease of variance is maximized.
- PCA results in components each of which contribute their part to the original data.
- The first principal component (PC) explains the most of original data variance, the second PC second most and so forth.
Principal Component Analysis (PCA) II

- If all PCs were recovered in the analysis, the original data could be represented exactly using PCs and converted data.
- The number PCs is the same as the original data dimension.
- Usually, far less components are utilized, which is the whole point of doing PCA on a data set.
- From pattern recognition viewpoint using PCA on measurement data, one is able to find out the most important features.
- Features that contribute the variance the least may be discarded (i.e., smaller feature vectors).
Hebbian based PCA I

- A single layer linear neural network is used to perform PCA.
- Properties
  - Each neuron in the output layer is linear
  - The network has $m$ inputs and $l$ outputs i.e. the number of elements in the original feature vector and the number of elements in the dimensionality-reduced feature vector ($l < m$).
- The synaptic weights of the single layer network are adapted according to Generalized Hebbian learning Algorithm (GHA)
Hebbian based PCA II

- Learning rule: \( w_i(n + 1) = w_i(n) + \eta y(n)x_i(n), i = 1, \ldots, m \)
  - \( w_i(n) \) denotes weight \( i \) in time step \( n \).\(^\dagger\)
  - \( x_i(n) \) is the input and \( y(n) \) is the output signal.
  - \( \eta \) is the learning rate.

- Note that the learning rule has the saturation problem (discussed earlier in this course). Solution: normalization of the update (See Haykin for details).

\(^\dagger\)The number of synaptic weights is equal to the number of inputs \( m \).
Questions 1

- Q1: What are committee machines?
- Q2: What are the basic structures of committee machines and what is their fundamental difference?
- Q3: Explain the fundamental idea of PCA.
- Q4: Consider a signal that is reconstructed using 10 and 100 principal components. What features both output signals have and what are characteristic to only one?
Bivariate data

Figure: Measurement data
Q5: What is direction of the first principal component (vector) for the measurement data in Figure 1‡.

‡The first component is the "first axis" in the new orthogonal basis.