5) How do neural networks learn?

• Iterative algorithm: weights of neural network are adjusted on-line as training data is received.
  \[ w(k+1) = L(w(k), x(k), d(k)) \]
  for supervised learning where \( d(k) \) is desired output.

• Need cost criterion: common cost criterion
  Mean Squared Error: for one output
  \[ J(w) = \sum (y(k) - d(k))^2 \]

• Goal is to find minimum \( J(w) \) over all possible \( w \). Iterative techniques often use gradient descent approaches.
Learning and Generalization

Learning algorithm takes training examples as inputs and produces concept, pattern or function to be learned.

How good is learning algorithm? Generalization ability measures how well learning algorithm performs.

Sufficient number of training examples. (LLN, typical sequences)

Occam’s razor: “simplest explanation is the best”.

Regression problem
Learning and Information (continued)

Generalization error

\[ J_g = J_{emp} + J_{model} \]

Empirical error: average error from training data (desired output vs. actual output)

Model error: due to dimensionality of class of functions or patterns

Desire class to be large enough so that empirical error is small and small enough so that model error is small.
Designing a pattern recognition system

Given a set of training data, design a system that can realize the desired task.
Assume two classes, binary outputs, input space $\mathcal{X}$ partitioned into two sets $\mathcal{X}^-$ and $\mathcal{X}^+$ that are disjoint and mutually exclusive.

- $x \in \mathcal{X}^+$, output $y=1$, represented by $+$
- $x \in \mathcal{X}^-$, output $y=-1$, represented by $\circ$
Gaussian data drawn from two classes
Classifier output

- Function output $y$ is binary valued.
- Learning algorithm takes labeled training examples and updates parameters $w$ so that most if not all training examples are correctly labeled.
- Pattern classification problems have different degrees of difficulty requiring appropriate binary function to perform classification.
Linearly separable

Set of points is linearly separable if a linear hyperplane can partition the + points from the o points.
Not linearly separable

A set of labeled points that cannot be partitioned by a linear hyperplane is not linearly separable.

Set of points that are not linearly separable
Linear threshold units

\[ sgn(s) = \begin{cases} 
1, & \text{if } s \geq 0 \\
-1, & \text{if } s < 0 
\end{cases} \]
Perceptron Learning Algorithm

An iterative learning algorithm that can find linear threshold function to partition linearly separable set of points. Assume zero threshold value.

1) \( w(0) = \text{arbitrary}, j=1, k=0 \)
2) Pick point \((x(j),d(j))\).
3) If \( w(k)^T x(k)d(k) > 0 \) go to 5)
4) \( w(k+1) = w(k) + \mu x(k)d(k), k=k+1 \)
5) Increment \( j \), if at end of data, set \( j=1 \), check if cycled through data without changing \( w \), if not go to 2
6) Otherwise stop.
PLA Example

\( w(1) \)
\[ x(1) = (-2, 1), d(1) = 1 \]
\[ x(2) = (-4, -1), d(2) = -1 \]
\[ x(3) = (1, -2), d(3) = -1 \]
\[ x(4) = (3, 1), d(4) = 1 \]
PLA comments

- Energy function
  \[ J(w) = - (\text{sum of synaptic strengths of misclassified points}) \]
  \[ w(k+1) = w(k) - \mu(k) \nabla J(w(k)) \]  (gradient descent)

- Margins
- Proof (Novikoff, requires margins)
- Homogeneously linearly separable
- Version space (weight space where feasible solutions lie)