# Evolutionary Strategies

Chromosom/Individuum -> realwertiger Vektor  
Gen -> Element des Vektors

**(1+1)-ES:**

Ein Elternteil -> (klonen) -> Ein Kind -> (Mutation) -> mutiertes Kind -> (Selektion) -> neue Population

P={p} c={c} c‘={c‘}

Weiterlaufen bis entweder vorgegebene Anzahl Generationen oder gewünschte Fitness erreicht wurde.

Mutation: zufällig normalverteilt mit Mittelwert = 0 und einer spezifischen Standardabweichung. Es gibt eine ideale Abweichung, die eine maximale Evolutionsgeschwindigkeit bringt. Zu wenig Varianz: Zu wenig Veränderungen. Zu viel Varianz: Zu große Änderungen führen zu nicht sinnvollen Mutationen.

**(µ+λ)-ES**

Nutzung von µ Eltern und λ Kindern, Auswahl aus der Union von Kindern und Eltern  
-> Eltern können mitgeschleppt werden (Unsterblichkeit), anfällig für lokale Maxima/Minima

**(µ, λ)-ES**

Wie oben, aber Auswahl der neuen Generation NUR aus Kindern, weniger anfällig für lokale Minima/Maxima

Selektionsdruck immer über das Verhältnis von µ und λ: ratio r=µ/ λ, 0 < r <= 1

**(µ/ρ # λ)-ES**

# steht für entweder + oder , und beschreibt, aus welcher Menge die neue Generation ausgewählt wird (und später auch, welche Generationen als neue Elterngenerationen gewählt werden können.

Kein Klonen mehr, Bildung von Paaren mit ρ Eltern und Rekombination der Gene via Mittelwert oder Crossover.  
Px=[3,7,2.8,9,5]  
Py=[7,7,2.0,5,1]  
r= [1,1,0 ,1,0] (Zufällig generiert bei Crossover)  
c1=[7,7,2.8,5,5]  
c2=[3,7,2.0,9,1]

**(m/r # l (µ/ρ # λ)/n)-ES**

Mehrere Parent-Populationen, Bildung von l Kinderpopulationen aus jeweils r gemischten Parent-Populationen. Auswahl der neuen Elternpopulationen aus den besten Kinder-Populationen (Im Falle `,`) oder aus Eltern- und Kind-Generationen.

# Genetische Algorithmen

Grundsätzlich:

* Alle können sich vermehren, nur Unterschied in der Wahrscheinlichkeit
* (Meist) Binäre Codierung der Gene
* Gene können entweder einzelne Bestandteile des Vektors oder auch mehrere Bestandteile zusammen sein.

Codierung:

Problem mit den Sprüngen im ASCII, deshalb verwendet man Gray-Code

### Schema Theorem und Building Block Hypothesis

Schema: Zeichenkette, die neben 1 und 0 noch beliebige Platzhalter enthält (-> Templatefunktion). Definiert das Set von Chromosomen, bei dem alle definierten Bits übereinstimmen.

Order of Scheme: Anzahl der definierten Bits

Definierte Länge: Abstand zwischen dem ersten und letzten definierten Bit

Anzahl der möglichen Schemata mit Länge L: 3L-1  
Anzahl der möglichen Chromosomen mit L: 2L

Auswirkungen der Selektion:

f(xi,t) Fitness von Chromosom xi zum Zeitpunkt t  
m(xi,t) Anzahl der Kopien des Chromosoms zum Zeitpunkt t  
p(xi)=Fitness/durchschnittliche\_Fitness Wahrscheinlichkeit, dass das Chromosom gewählt wird

Fitness eines Schemas: Nach der Anzahl der Kopien gewichteter Mittelwert der Fitness aller Chromosomen eines Sets

# Neuronale Netzwerke

* Algorithmen zur Verarbeitung von Daten, deren Prinzipien von biologischen Nervenzellen abgeschaut wurden.
* Ein mathematisches Neuron ist charakterisiert durch eine Eingabedimension n, einen Gewichtsvektor w und eine Aktivierungsfunktion f sowie einen Schwellwert/Bias Θ
* Man unterscheidet (im Wesentlichen) zwischen 2 Neuronenarten:
  + Skalarproduktbasierte Neuronen, bei denen die Funktion eine Summation aller Inputs und deren Gewichte ist (Also praktisch das Skalarprodukt aus w und Eingabevektor x transponiert)
  + Metrik-/Normbasiert: Funktion ist die Summe der Differenzen von Eingabe x und Wichtung w
* Wichtige Typen:
  + Perzeptron: x -> H(<x,w> - Θ) [Pyramidenzelle, skalarproduktbasiert)
  + Sigmoidaler Neuron:
  + Radiale-Basis-Funktion-Neuron:
* Ein Neuronales Netz ist ein Tupel (N,->,w, Θ,f,I,O)
  + N… Neuronen (Index der …)
  + -> … NxN – Verbindungen zw. Den Neuronen
  + w… Gewichtsvektor (wi,j) der Verbindungen
  + Θ… Schwellwerte
  + f… Vektor der Aktivierungsfunktionen
  + I… Eingabeneuronen (Inputs)
  + O… Ausgabeneutronen (Outputs)
* Neuronale Architektur: Ein neuronales Netz, bei dem Gewichte und Schwellenwerte noch nicht festgelegt sind, sondern Variablen darstellen
* Vernetzungsstrukturen:
  + Feed-Forward Network (FFN): (N,->) ist ein azyklischer Graph, bei denen zwischen I und O klar getrennt wird
    - Trennung der Neuronen N in verschiedene Layer (1 (Input) bis h (Output)), innerhalb dieser Layer kann nicht vernetzt werden, nur zu anderen Layern.
    - Alle Layer zwischen I und O als hidden Layer zusammengefasst, hidden Layer nicht zwingend
    - Verbindungen über mehrere Layer hinweg werden shortcuts genannt
  + Voll Vernetztes Multilayer FFN:
    - Sonderfall des FFN, Jede Schicht ist vollständig mit den Neuronen der nächsten Schicht vernetzt, ohne shortcuts
  + Rekurente Netze (vollst. Vernetzt)
    - -> NxN, also I=O=N
* Aktivierungsfunktionen
  + F=H, d.h. f= Summation der Gewichte mal der Inputwerte (Hearisidefunktion)
  + Bipolare Aktivierung: 1 für x>=0, 0 für x<0
  + Id(z)=z (Identische Funktion)
  + Semi-linear (=Input, Cutoff für Input > 1 oder < 0)
  + Sigmoid: sgd(z)=1/(1+exp(-z))
  + tanh(z)=(ez-e-z)/(ez+e-z)
  + Relu: Rectified linear with funktion (=Input, aber Cutoff für Input < 0)

Lernen in neuronalen Architekturen bedeutet Anpassung der Gewichte und Schwellwerte, um eine Lernaufgabe zu erfüllen.

Lernalgorithmus:

* Initialisierung mit zufälligen Gewichten und Schwellenwerten
* Solange mindestens ein Datenpunkt falsch klassifiziert wird: Anpassung der Gewichte in Richtung des Fehlers
* Der PLA konvergiert in endlich vielen Schritten, falls die Lernaufgabe lösbar ist.
* Eine Lernaufgabe heißt lösbar für ein Perzeptron, solange die Datenpunkte der beiden Klassen linear trennbar sind.
* Problem: Lösbarkeit des Problems ist nicht vorher bekannt und kann zu unendlich langen Durchläufen führen.
  + Heuristischer Ansatz: Abbruch zu einer Zeit k1
  + Dann Teilung der Datenmenge in richtig und falsch klassifiziert
  + Training eines neuen Perzeptrons mit der falsch klassifizierten Datenmenge
  + Wiederholung nach Bedarf
  + Schwierigkeit: Wann soll abgebrochen werden? Wie kann bestimmt werden, ob das erste Perzeptron schon gut gelernt hat?
  + Perzeptron-Pocket-Alg.: Für optimales Perzeptron im nicht-trennbaren Fall
    - Nach jeder Änderung der Gewichte Prüfung, wie viele Datenpunkte richtig klassifiziert wurden bzw. wie wenig Fehler gemacht worden und Speicherung der besten Gewichte
* Ansatz für Mehrklassenproblem:
  + Ebenfalls Heuristik
  + Auswahl einer Klasse gegen alle anderen Klassen und lernen der einzelnen Klasse
  + Dann neues Perzeptron für nächste Klasse
  + -> Kaskade von Perzeptrons

Lineare Perzeptrons

Aktivierungsfunktion linear, nicht mehr Heariside/Stufenfunktion

* Fehlerfunktion: SSE (Sum Squared Error): Mittelwert der quadrierten Abweichungen der Klassenlabel (y) von dem erhaltenen Funktionswert
* Bestimmung der Gewichte so, dass SSE minimal ist
  + -> Ableitung der Fehlerfunktion (SSE) nach den Gewichten = 0 an diesem Minimalpunkt
  + Trick: Bestimmung der lokalen Fehler aller Punkte und Summation dieser.
  + Le(xp,w)=(wTxp-yp)xp
  + (Updateschritt)
* SGDL für SSE: mit random(xP, yP)

## Feed-Forward-Netze

* Nutzung von sigmoiden Perzeptrons
  + Kann an Heariside-Funktion angenähert werden
  + Unendlich ableitbar
* Wie viele Hidden Layer? Wie viele Neuronen pro Layer?
  + Für lineare Perzeptronen nur eine hidden layer, da sich aus der Kombination von linearen Funktionen nur lineare Funktionen ergeben können
  + Theorem P. Werbos: Das Multilayer Perzeptron kann eine beliebige Funktionen f im reelen Raum beliebig genau approximieren, falls f mindestens 2 mal differenzierbar ist. Dafür reicht eine hidden layer. (Nur gültig, wenn Funktion nicht linear, differenzbar und beschränkt ist.
  + Anzahl der benötigten Neuronen im hidden Layer unbekannt und nicht vorher bestimmbar
  + Ausweg: Trainieren eines sehr großen Netzes, Berechnung der SSE und pruning von Neuronen, solange sich dadurch der SSE noch verbessert
  + Problem bei Netzen mit vielen Schichten: Nur die letzten layer vor dem Output werden ordentlich trainiert, frühere Gewichte kaum betroffen (Vanishing gradient)
    - Anstelle der sigmoidalen Funktion vielleicht lieber als Annäherung mit Relu arbeiten, leichter (schneller) ableitbar
    - Blockprinzip nach Hinton: Zerlegung des Netzwerkes in viele einzelblöcke aus input und „bottle neck“ (den verbundenen Neuronen), das einen Output der gleichen Dimension erzeugt, die einzeln optimiert werden. Genannt Autoencoder-Netzwerke

## Deep Networks

* Multilayerperzeptron mit vielen Hidden Layern
* Üblicherweise nach dem Blockprinzip von Hinton (s. ein paar Zeile weiter oben) trainiert
* Beispiel: **Bilderkennung/Klassifikation**
* Jede Schicht ist eine Art Filter, die die Daten transformieren, z. B. Faltung oder Pooling der Daten
* Am Ende entstehen Datenblöcke, deren räumlicher Zusammenhang (Reihenfolge der Eingabe der Inputs) absichtlich ignoriert wird
* Voraussetzung für gutes Deep Network:
  + Gut vortrainiertes Netz (z.B. google, Amazon)
  + Oder Extreeeem viele Daten

# c-Means

//aufzeichnungen fehlen xD

## Neural Gas

Neural Gas: Verbessertes k-Means mit „Abkühlfaktor“ λ  
Gegeben ist eine Datendichte P(v). Erwünscht wäre also eine Dichte der Prototypen äquivalent zur Datendichteα (Viele Prototypen im Bereich vieler Datenpunkte, α=1 wäre das Informationsoptimum)  
Theorem: deff sei die effektive Dimension der Daten. Für 0 < λ << Anzahl der Prototypen im NG, gilt nach dem Training: . Demzufolge werden hochdimensionale Daten optimaler dargestellt als niederdimensionale.

T. Kohone

Cortical Area – 2D  
2d-grid of Neurons r = (r1, r2)  
Jedes Neuron ist eine Pyramidenzelle: Or=wr\*v (Erregung = Gewichtsvektor mal Reiz)  
Maximale Erregung: max(Or) = min(v-wr)² über alle r Element A  
=> Sieger S(v)=argminr(v-wr)²  
Gebiet der maximalen Erregung kann auf dem grid dargestellt werden.  
Bei Erregung nach Hebbscher Lernregel Anpassung der Gewichte: ΔwS(v)=ε(v-wS(v))  
Es fehlt noch die Nachbarschaftskooperation: Auch benachbarte Neuronen wollen mitfeuern und sich anpassen. => Δwr=ε hσ(r,s)(v-wr) mit hσ(r,s)= (Nachbarschaft zwischen r und Sieger s in A) *Unterschied zum Neural Gas: Abstandsfunktion nicht nach Rangfolge der Seigerfunktion sondern nach Nachbarschaft*

Aber: Man kann zeigen, dass es keine Kostenfunktion gibt, die mit dieser Siegerfunktion und dieser Lernregel optimiert wird.

Deshalb Heskes: S(v)=argminr(Summe über alle anderen r‘(hσ(r,r‘)\*(v-wr)²)) und Δwr=ε hσ(r,s(v))(v-wr)

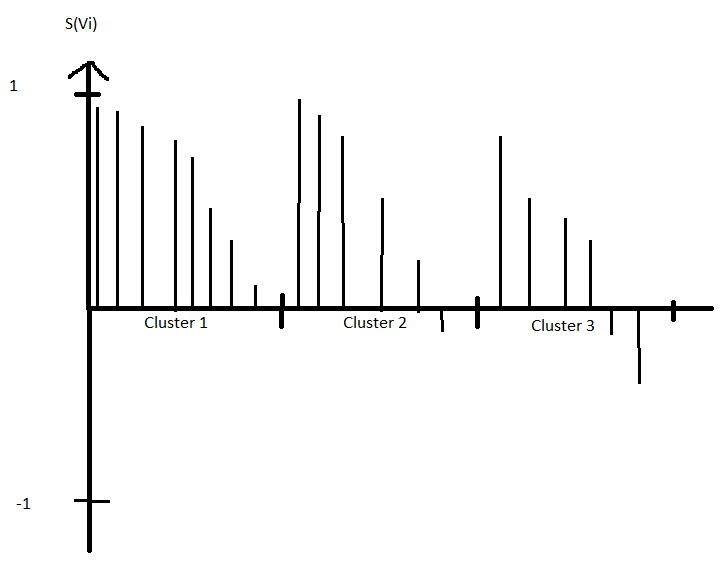
Silhouette Analyse

* Graphische Reräsentation der (harten) Cluster
* Indexwert misst Ähnlichekit des Datenpunkts zum eigenen Cluster gegen die Ähnlichkeit zum nächstgelegenen Cluster

S(Vi)… Silhouetten-Index des Datenpunkts

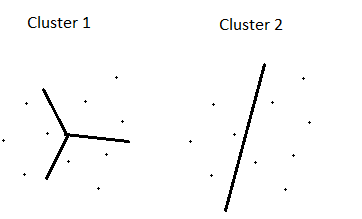
S(Vi)= 0 wenn identisch dem eigenen Cluster und dem nächstgelegenen Cluster  
Ansonsten: ((Abstand zum eigenen Cluster) – (Abstand zum nächsten Cluster))/max(AeigenesCLuster, Anächstes Cluster)  
Abstand eines Punktes zum Cluster: Mittelwert des Abstandes des Punktes zu allen Punkten des Clusters  
-1 <= S(Vi) <= 1   
=> etwa 0: liegt zwischen beiden Clustern  
S(Vi)<0: Liegt näher an anderem Cluster  
S(Vi)>0: näher am eigenen Cluster

Silhouetten-Index des Clusters:  
S(Ωi) ist der Mittelwert der Silhouetten-Indizes der Datenpunkte des Clusters  
Silhouetten-Index des Datensatzes:  
S ist der Mittelwert aller Silhouetten-Indizes der Datenpunkte

Graphische Darstellung: Nach Clustern getrennt und nach Datenpunktindex sortiert aufgetragen.  


Rand Index

* Misst Ähnlichkeiten zwischen 2 Clusteraufteilungen
* Paarweiser Vergleich: zählt Übereinstimmungen und Nicht-Übereinstimmungen
* Anzahl der Cluster muss nicht übereinstimmen

Jedes mögliche Paar aus Punkten wird betrachtet, ob sie beiden Punkte in Lösung 1 und Lösung 2 in einem Cluster zusammengefasst wurden oder nicht.

|  |  |  |  |
| --- | --- | --- | --- |
| Fall | C1 | C2 | Anzahl |
| A | Zusammen | Zusammen | … |
| B | getrennt | Zusammen | … |
| C | Zusammen | Getrennt | … |
| D | Getrennt | Getrennt | … |

RI=(A+D)/(A+B+C+D)  
0<=RI<=1  
RI=0 => keine Übereinstimmungen  
RI=1 => komplette Übereinstimmung

Kappa Index

* Statistisches Maß, misst Übereinstimmungen zwischen 2 oder mehr Klassifikationen
* Unter bestimmten Voraussetzungen auch zur Clusterevaluierung einsetzbar:
  + Cluster müssen eindeutig identifizierbar und zuordenbar sein
  + Anzahl der Cluster muss übereinstimmen

K=(Fehler durch Algorithmus)/(Fehler durch vollständige zufällige Übereinstimmung)

K in[-1,1]  
K<0 => Schlechte Übereinstimmung  
0,8<=K<=1 => (fast) perfekte Übereinstimmung

Machine Learning

|  |  |  |
| --- | --- | --- |
| Unüberwacht | Überwacht | Reinforcement |
| k-means, NG, SOM | Multilayer Perceptron |  |
| Analysieren  Visualisieren  Clustern | Klassifikation  Regression |  |

# Klassifikation

Teilung der Datenmenge:

|  |  |  |
| --- | --- | --- |
| Training | Test | Validierung |
| Erstellen des Modells | Ermitteln der besten Parameter (werden nicht optimiert) | Evaluation des Modells |

n-fache Kreuzvalidierung:  
Zerlegung in n gleich große, disjunkte Teilmengen.  
Jede Teilmenge 1 mal für Test, den Rest für Training