Neural Networks forelesning 4

* Hva slags input-funksjoner
  + Hvorfor ikke noes om er litt glattere enn en lineær funksjon
  + Hvorfor bruke de man har
  + Hvordan bør man initialisere vektene sine
  + En lineær separator klarer vare å separere lineære tasks=> Grunn for sigmoid
  + Hva er gode modeller for regularisering, og hvorfor
  + Man kan initialisere med å trekke ut et subset for hvert av disse modellene
  + Vi kan ta gjennomsnittet av disse
  + Hver modell har sin versjon av datasettet
    - Gjennomsnittet skal visstnok alltiv være bedre
    - Bagged output:
      1. Utsagn, den baggede outputen er alltid bedre enn enkelt-outputsene
      2. ….
      3. Bagged error <= Normal error

Del 2

* Dropout:
  + Kjent oppsett for kunstige neurale nettverk
  + Vi blokker enkelte neuroner. Men for eksempel halvparten av neuronene
    - Forskjellig konfigurasjon for hver gang
    - Når man tester ut nettverket bruker man hele nettet
    - Nettverket må også skalleres for at det skal fortsette å være riktig
    - Bagging brukes ofte som en motivasjon for hvorfor dropout virker
  + Dropout som tvungen feature sharing
    - Det tar dobbelt så mange neuroner, men det sørger for at alle løser liknende oppgaver
      1. Gir mer «Robuste» resultater
      2. Mandre enkle klassifiseringsmetoder
         * For bildebehandling kan man gjøre en del ekstra ting

Støy, Hue shift, elastisk deformering, tilfeldig translasjon

Om man kjenner til en del av de intrinsike symmetriene i et problem kan man lage flere samples fra en enkel en

Man kan gjøre transformasjoner som man vet at hører til i samme klasse

* + - 1. Early stopping (Egentlig mer «Retrun to best previous”)
      2. Smart vekt-initialisering
    - Initialisering (Input-laget)
      1. Hvordan bør man velge variansen
      2. Claim: roten av n er viktig

Initialisering

* Hvorfor er initialisering oller normalisering i backprop
* Forward pass: Linear og nonlinear processing
  + En lineær funksjon er uinteresant, fordi hvert lag kan sees på som en multiplikasjon med matriser. (Det er et problem med matriser, man ender opp med en matrise til slutt)
  + Man får bare interesante egenskaper om man har forskjellige slopes, og brudd-punktet
    - Om alle nettverket rører den samme lineære path-funksjonene, da er hele nettverket lineært
    - Forskjellige mønster bør røre ved forskjellige regioner
  + Initialiseringen må være slik at forventningsverdi er 1 og standardavvik er av orden 1
  + Vanishing gradient problem
    - Hve path bidrar....(?)
* Obsercasjoner
  + For hver enkelt-path er det mange termer g’
* Konklusjon
  + Forward pass:
    - Må unngå linearitets-problemet
  + Backward pass
    - Må unngå vanishing gradient problemet(?)
* Weight update step
  + ReLu
  + Tanh
  + Shifted exponential Linear (SEUL) vs tanh
    - 1. Problem for ReLu og andre enheter med non-negative x
      2. Ikke et problem for tanh eler SELU
      3. Men man får vanishing gradient, …
  + Selu løser
    - Vanishing gradient
    - Biasing output
    - …. En sitste
  + Om man venter lenge nok, så er det rellativt likt
    - Initialisering er viktig for å unngå linearitets-problemet.