Neuronale Netze (SS 2002), 15.5.

A short look at probability theory:

- A **probability room** is a set of possible events (a σ-algebra) together with a probability measure P such that the probability of every (measurable) subset of the set can be measured by P.
  - $P \geq 0$, $P(\emptyset) = 0$,
  - $P(A^c) = 1 - P(A)$, $P(\bigsqcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$

  In our case:
  - discrete set $\{x_1, x_2, \ldots\}$ such that $P(x_i)$ defines P
  - or continuous set $\mathbb{R}^n$ such that the probability is given by a density function p: $P(A) = \int_A p(\vec{x}) d\vec{x}$.
- **random variable** X is a (measurable) function from a probability room to the real numbers
- **expectation** $E(X) = \lim_{n\to\infty} \sum X(x_i)/n$ (x_i chosen according to the given probability) is the expected value of X (if existing)
  - for discrete sets: $E(X) = \sum P(x_i)X(x_i)$ if existing
  - for continuous sets: $E(X) = \int p(\vec{x})X(\vec{x})d\vec{x}$ if existing
  - $E(X + Y) = EX + EY$, $E(\lambda X) = \lambda EX$
- **variance** $Var(X) = E((X - EX)^2)$ (if existing) measures how the several events differ from the expected value, for a constant this is 0
  - $Var(X + \lambda) = VarX$, $Var(\lambda X) = \lambda^2 VarX$,
  - $VarX = E(X^2) - (EX)^2$
  - $VarX = \lim_{n\to\infty} \sum (x_i - \sum x_j/n)^2/(n-1)$ if x_i is chosen according to the given probability

Bias-Variance dilemma of architecture selection:

a set of data $(\vec{x}_i, f(\vec{x}_i) + \eta)$, where η is noise, f is to be learned.

- a small architecture cannot fit the data; but it will not learn the noise, i.e. the behavior generalizes to unseen data
- a large architecture can fit data; but it will fit the noise, too, i.e. it does not generalize to unseen data
- mathematics: η depends on the underlying distribution P, assume f_w is the network output

  $$E_P((f_w(\vec{x}) - f(\vec{x}))^2) = (E_P(f_w(\vec{x})) - f(\vec{x}))^2 + E_P((f_w(\vec{x}) - E_P(f_w(\vec{x})))^2)$$

  i.e. the quadratice error divides into the bias, i.e. the deviation of the expected outcome from the real one, and the variance, i.e. the deviation of the outcome from the expected outcome in several runs

Mathematical guarantees:

- For every $\epsilon > 0$, compact set C, and every continuous function $f : C \to [0, 1]$ a fully connected sigmoidal feedforward network with one hidden layer f_w can be found such that

  $$|f(\vec{x}) - f_w(\vec{x})| < \epsilon \forall \vec{x} \in C$$

- For every finite set of m points $(x_i, y_i) \in \mathbb{R}^n \times]0, 1[$ which is not contradictory, a fully conneted sigmoidal feedforward network with one hidden layer with m hidden neurons can be found and

  $$f_w(\vec{x}_i) = y_i \forall \vec{x}_i$$