Course ID 80250993 Dates: 9/16-12/23/2021 @ I-205
Tencent/VooV Meeting 521 4678 6257 (passwd 1205)

Chapter 5
Classical Artificial Neural Networks (continued)

Xuegong Zhang
October 7, 2021

Basic concepts of ML: MLP

- How can we make a learning machine?
  - It needs a teacher.
    - → The model: $g(\mathbf{x}) = f\left(\sum_j w_{jk} f\left(\sum_i w_{ij} x_i + w_{j0}\right) + w_{k0}\right)$
  - We need materials to train it. / It needs materials to learn from.
    - → Training data: $\{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)\},\ \boldsymbol{x}_j \in R^{d+1}, y_j \in R$
  - We need to tell what is the goal of the learning.
    - → Objective function: $\min E = \frac{1}{2}\sum_{j=1}^{N}\left(g(\boldsymbol{x}_j) - y_j\right)^2$
  - We need to tell it how to learn.
    - → Learning algorithm: $\boldsymbol{w}(k+1) = \boldsymbol{w}(k) - \rho_k \nabla E$ via the BP algorithm

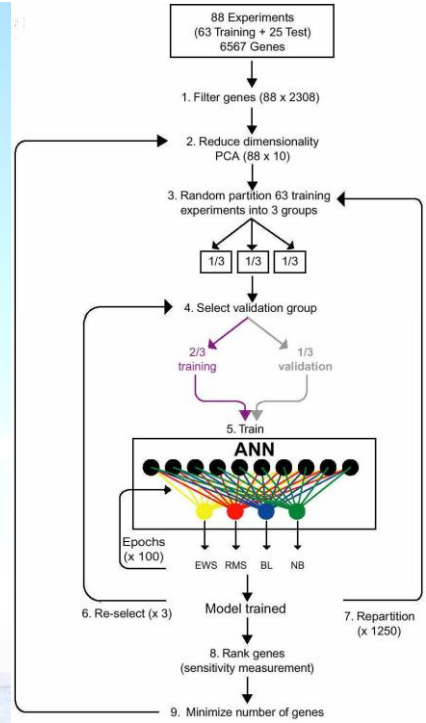
What can an MLP be used for?

D

Example:
ANN for cancer classification with gene expression data

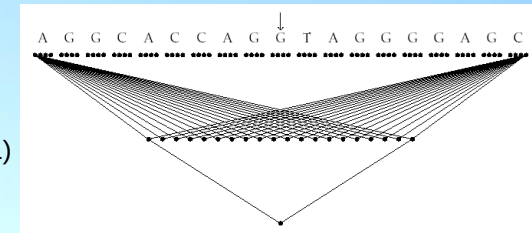
- Multi-class:
  4 classes, one output per class
- A set of NN models, then voting
- PCA for feature extraction

- refs.
  - Khan, J. … and Meltzer, P.S. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Medicine*. **7**(6): 673-679, 2001
  - Gruvberger, S. … and Meltzer, P.S. Estrogen receptor status in breast cancer is associated with remarkably distinct gene expression patterns. *Cancer Res.* **61**: 5979-5984, 2001



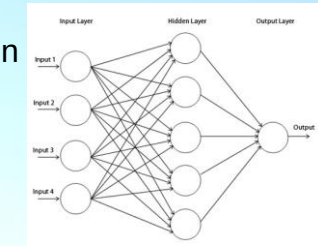
Example: MLP for splicing site prediction

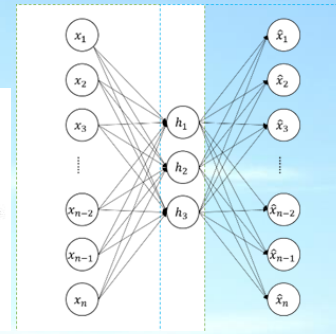
- 3-layer MLP, BP learning
- Input:
  - a DNA sequence segment of 15 nucleotides
  - encoded as A(1000), C(0100), G(0010), T(0001)
  - Total input dimension: 15x4=60
- Output: 1 node, value [0,1], threshold 0.5
- Similar applications:
  - protein secondary structure prediction
  - subcellular location prediction
  - …



What else can MLP do?

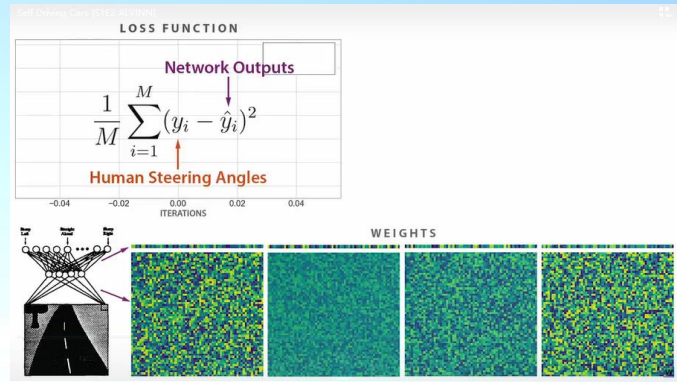
- Other applications
  - Nonlinear regression / function estimation
  - Time series prediction
  - Data compression / representation
  - …





Example: MLP for Autonomous Driving:
ALVINN and Autonomous Navigation in 1983-1995

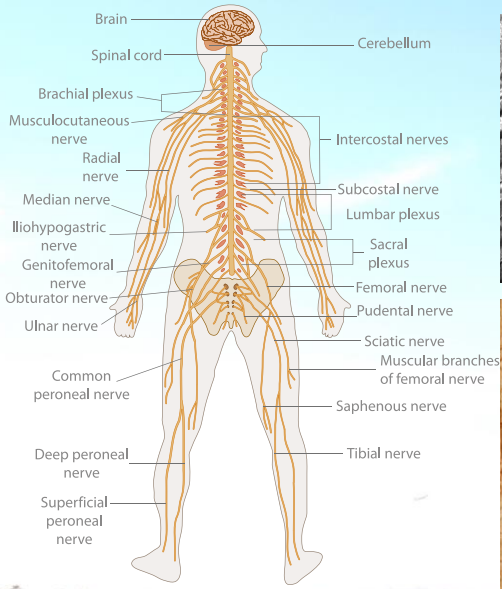
Stanford CS229 Lecture 2, 3:21-8:40

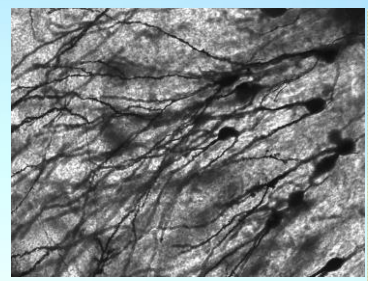


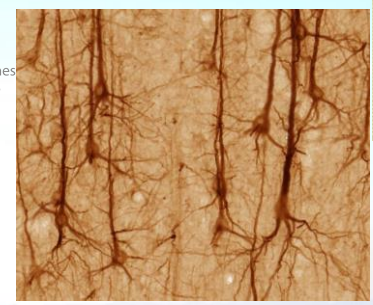
Self driving cars [S1E2 ALVINN].mp4 ~ 8+11min

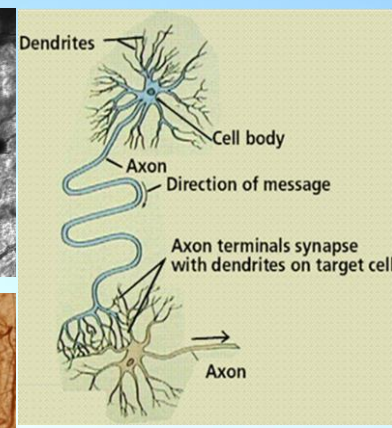
Why do we call it a “Neural Network”?

Neurons



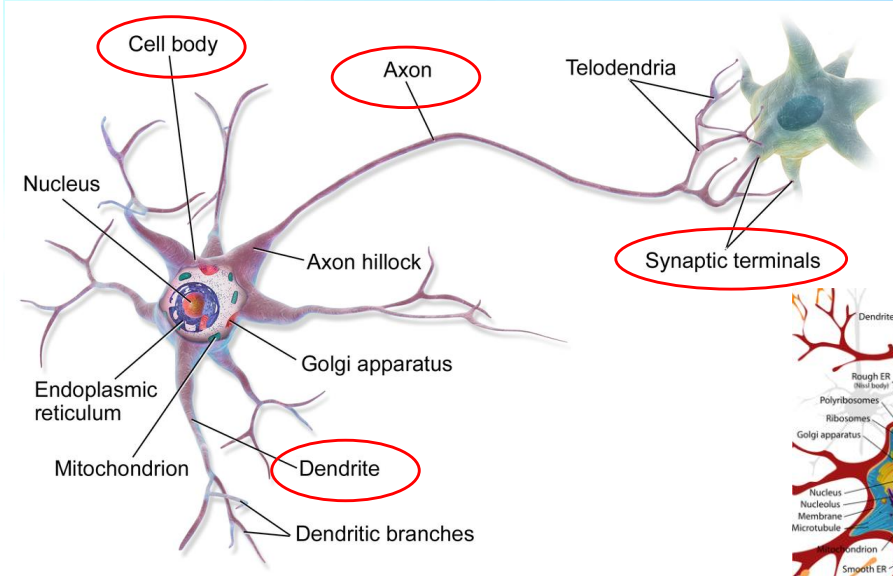




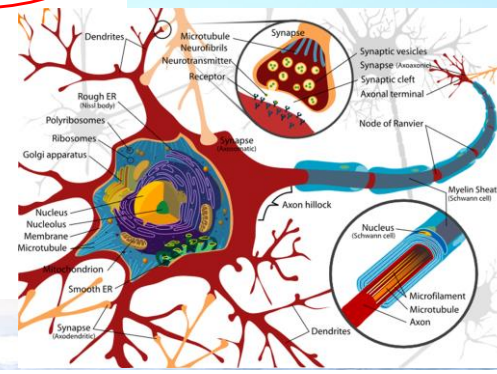


-- Wikipedia

A Neuron



- Neuron 神经元
  - cell body dendrites 树突
  - axon 轴突
  - synapses 突触



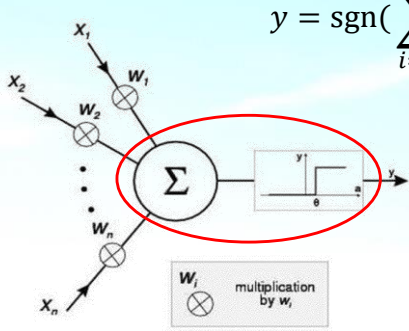
Xuegong Zhang

https://en.wikipedia.org/wiki/Neuron

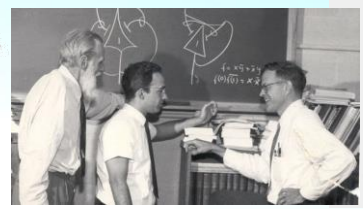
An artificial neuron

- A mathematical function as a model of a biological neuron
  - The linear threshold function (TLU- Threshold Logic Unit) first proposed by McCulloch and Pitts in 1943

$$y = \text{sgn}\left(\sum_{i=1}^{n} w_i x_i + w_0\right), \qquad y = \begin{cases} +1 & \Rightarrow \text{class } A \\ -1 & \Rightarrow \text{class } B \end{cases}$$



BULLETIN OF
MATHEMATICAL BIOPHYSICS
VOLUME 5, 1943



A LOGICAL CALCULUS OF THE
IDEAS IMMANENT IN NERVOUS ACTIVITY

WARREN S. MCCULLOCH AND WALTER PITTS

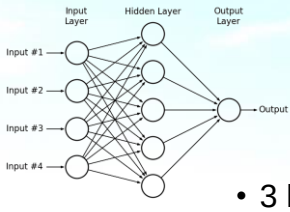
FROM THE UNIVERSITY OF ILLINOIS, COLLEGE OF MEDICINE,
DEPARTMENT OF PSYCHIATRY AT THE ILLINOIS NEUROPSYCHIATRIC INSTITUTE,
AND THE UNIVERSITY OF CHICAGO

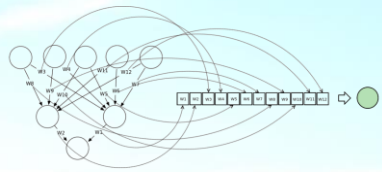
Xuegong Zhang

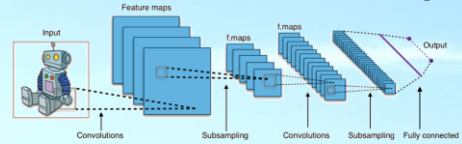
10

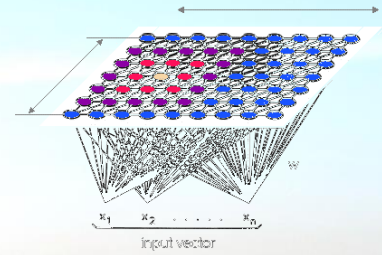
General structure of ANNs

- A large amount of simple computing units (nodes)
- connected as a network to perform a certain function
- with parameters like the connecting weights learnable from data according to a learning algorithm







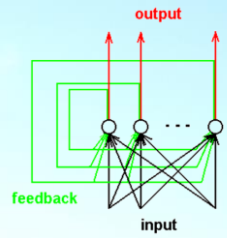


- 3 key aspects：
  - The connection structure
  - The activation function of neuron
  - The learning algorithm

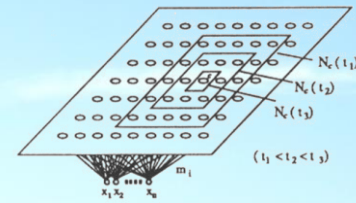
Three Major Types of ANN (in 1980s)

- Feedforward NN
  e.g., Multi-Layer Perceptron (MLP)

- Feedback NN
  e.g., Hopfield Net



- Competitive learning NN
  e.g., Self-Organizing Map (SOM)



Related statistical methods

- MLP

$$g_k(\mathbf{x}) \equiv y_k = f\left(\sum_j w_{jk} f\left(\sum_i w_{ij} x_i + w_{j0}\right) + w_{k0}\right)$$

- Projection pursuit

$$z = \sum_{j=1}^{j_{max}} \boldsymbol{w}_j f_j\left(\boldsymbol{v}_j^T \boldsymbol{x} + u_{j0}\right) + w_0$$

- Generalized additive model

$$z = f\left(\sum_{i=1}^{d} f_i(\boldsymbol{x}_i) + w_0\right)$$

- Multivariate adaptive regression spline (MARS)

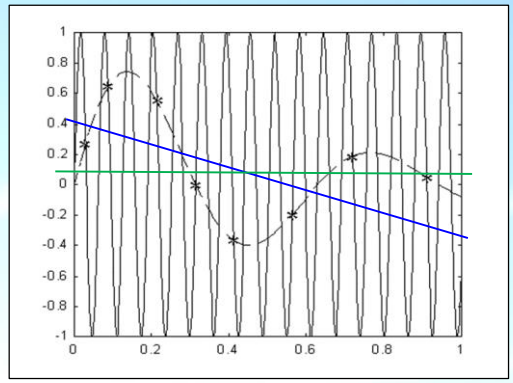
$$z = \sum_{k=1}^{M} \boldsymbol{w}_k \prod_{r=1}^{r_k} \phi_{kr}\left(x_{q(k,r)}\right) + w_0$$

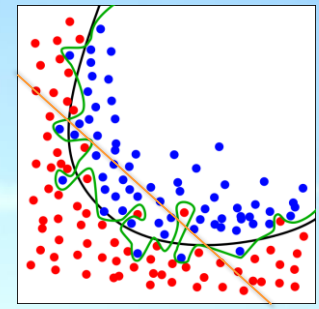
Duda, Hart & Stork, Pattern Classification, John Wiley & Sons, 2001, p.305-306

Discussions:
MLP structure design and algorithm tuning

Generalization of LMs

- Under-fitting
- **Over-fitting**





https://en.wikipedia.org/wiki/Overfitting

The learning curve and the use of validation set

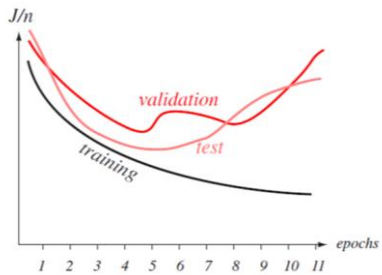


FIGURE 6.6. A learning curve shows the criterion function as a function of the amount of training, typically indicated by the number of epochs or presentations of the full training set. We plot the average error per pattern, that is, $1/n \sum_{p=1}^{n} J_p$. The validation error and the test or generalization error per pattern are virtually always higher than the training error. In some protocols, training is stopped at the first minimum of the validation set. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Observations:

- Test error does not always decrease with training error
- Using validation set to find possible signs of overfitting during training
- Sometimes training error itself does not decrease monotonically

Error surfaces of MLP

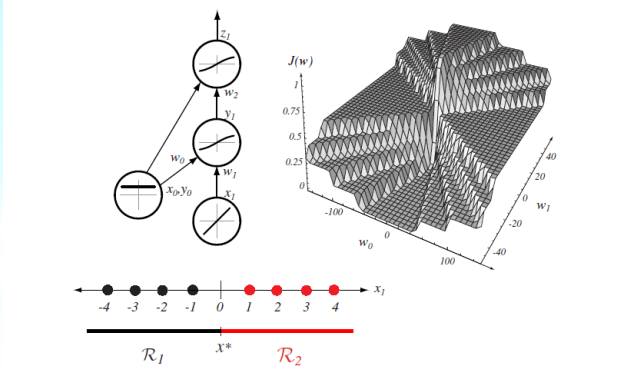


FIGURE 6.7. Eight one-dimensional patterns (four in each of two classes) are to be learned by a 1-1-1 network with steep sigmoidal hidden and output units with bias. The error surface as a function of w_0 and w_1 is also shown, where the bias weights are assigned their final values. The network starts with random weights; through stochastic training, it descends to a global minimum in error. Note especially that a low error solution exists, which indeed leads to a decision boundary separating the training points into their two categories. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

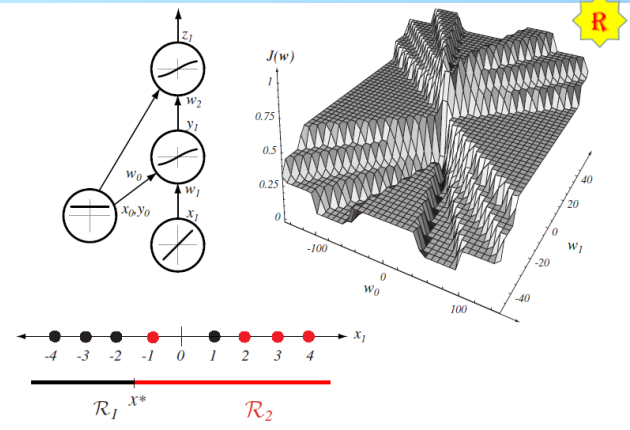


FIGURE 6.8. As in Fig. 6.7, except here the patterns are not linearly separable; the error surface is slightly higher than in that figure. Note too from the error surface that there are two forms of minimum error solution; these correspond to $-2 < x^* < -1$ and $1 < x^* < 2$, in which one pattern is misclassified. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Error surfaces of MLP

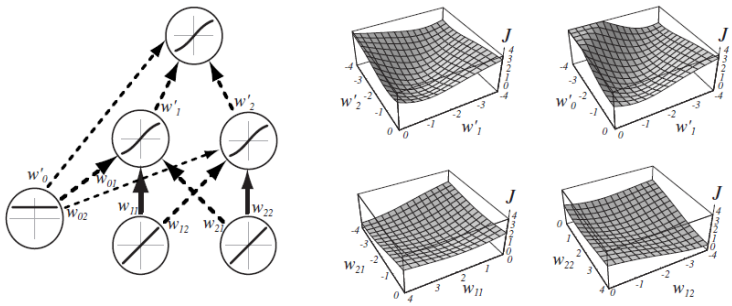


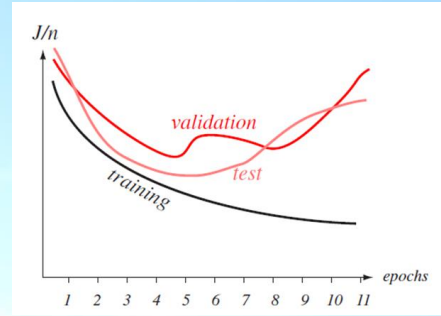
FIGURE 6.9. Two-dimensional slices through the nine-dimensional error surface after extensive training for a 2-2-1 network solving the XOR problem. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Discussion:

- Error surfaces of larger networks (higher dimensional parameter space) are more complicated
- The error varies gradually with regard to a single weight
- May afford more ways (dimensions) not to be trapped in local minima

Problems often met in BP learning

- Slow convergence
- Oscillating learning curve
- Overfitting: Poor generalization
- Possible reasons
  - Improper structure or activation function
  - Insufficient or improper training samples
  - Improper initial values, learning rates, …
  - Improper pre-processing
  - Need better training strategy
  - Bad luck
  - Poor problem setting



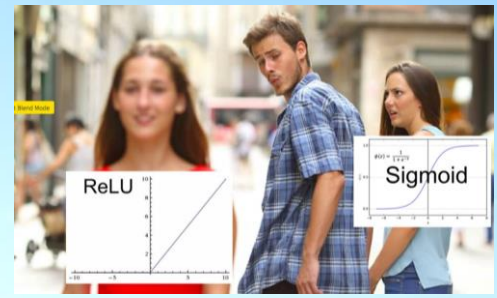
Popular protocols of BP Training

- Stochastic training (SGD)
  - Training samples chosen randomly, and weights updated with each sample
- Batch training
  - All training samples presented before learning takes place → an epoch
- Mini-batch training
  - Each epoch uses a group of samples for the training
- On-line training
  - Each sample is presented once and only once
- Learning with queries
  - The output of the NN is used to *select* new training samples
    → latter called “Active Learning”

Duda, Hart & Stork, Pattern Classification, John Wiley & Sons, 2001, p.293-295, 480-481

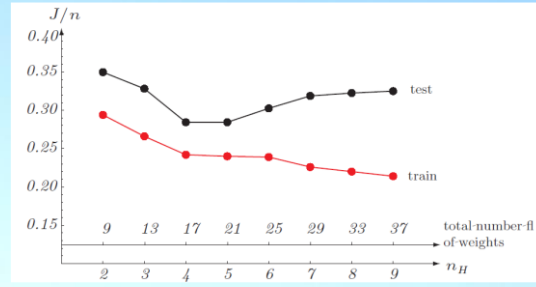
Practical techniques for improving BP

- Activation function (transfer function)
  - Sigmoid, tanh, ReLU, …
- Scaling input / standardization
  - effective region
  - relative importance
- Target values
  - use 0.1 vs. 0.9 instead of 0 vs. 1 for the two classes
- Initialization
  - Initialize in proper range to ensure *uniform learning*



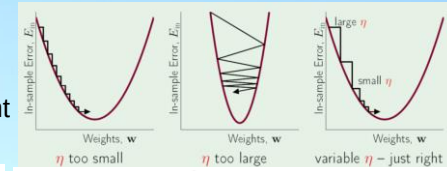
Practical techniques for improving BP

- Training with augmented pseudo samples
  - noisy samples
  - invariant “manufactured” samples
  - balancing samples
- Number of hidden units / number of layers
  - w.r.t. complexity of problem
  - ~n/10?
  - Pruning
  - Regularization



Practical techniques for improving BP

- Learning rates
  - $\eta_{opt} = (\partial^2 J/\partial w^2)^{-1}$ for quadratic objective function
  - Calculate the 2nd derivative with regard to *each* weight
  - Set optimal rate separately for each weight



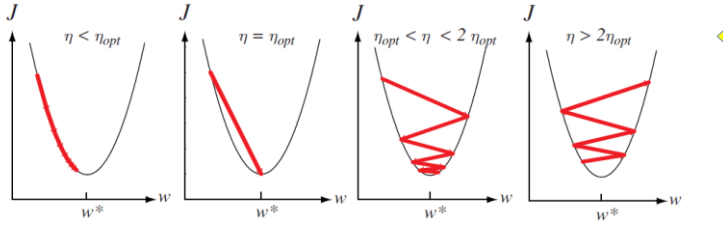


FIGURE 6.16. Gradient descent in a one-dimensional quadratic criterion with different learning rates. If $\eta < \eta_{opt}$, convergence is assured, but training can be needlessly slow. If $\eta = \eta_{opt}$, a single learning step suffices to find the error minimum. If $\eta_{opt} < \eta < 2\eta_{opt}$, the system will oscillate but nevertheless converge, but training is needlessly slow. If $\eta > 2\eta_{opt}$, the system diverges. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

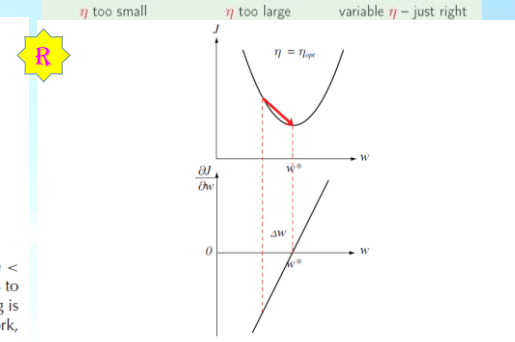


FIGURE 6.17. If the criterion function is quadratic (above), its derivative is linear (below). The optimal learning rate η_{opt} ensures that the weight value yielding minimum error, w^*, is found in a single learning step. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Practical techniques for improving BP

- Momentum

$$\boldsymbol{w}(t+1) = \boldsymbol{w}(t) + (1-\alpha)\Delta\boldsymbol{w}_{bp}(t) + \alpha(\boldsymbol{w}(t) - \boldsymbol{w}(t-1))$$

- Weight decay:

$$\boldsymbol{w}^{new} = \boldsymbol{w}^{old}(1-\epsilon)$$

  - Equivalent to the gradient descent of the new effective error with *a regularization term*

$$J_{reg}(\boldsymbol{w}) = J(\boldsymbol{w}) + \frac{2\epsilon}{\eta}\boldsymbol{w}^T\boldsymbol{w}$$

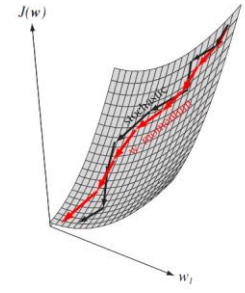
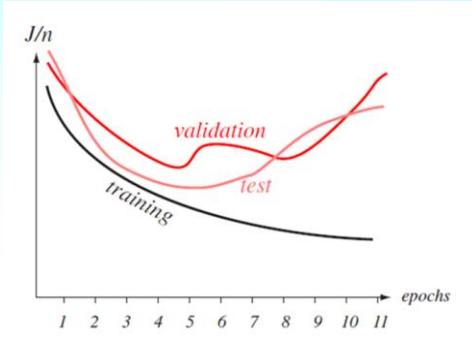


FIGURE 6.18. The incorporation of momentum into stochastic gradient descent by Eq. 37 (red arrows) reduces the variation in overall gradient directions and speeds learning. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Practical techniques for improving BP

- Stopping criteria
  - Avoid *overfitting*, stop before gradient vanishes, use a validation set to help
- Hints (augmented output with hint units) and Pretraining



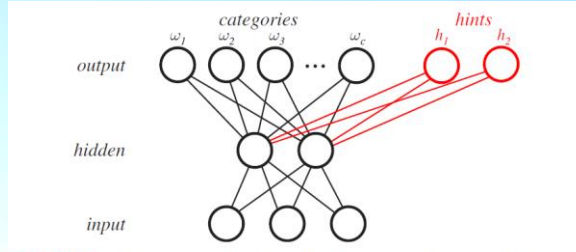


FIGURE 6.19. In learning with hints, the output layer of a standard network having c category units is augmented with hint units. During training, the target vectors are also augmented with signals for the hint units. In this way the input-to-hidden weights learn improved feature groupings. The hint units are not used during classification, and thus they and their hidden-to-output weights are removed from the trained network. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Duda, Hart & Stork, Pattern Classification, John Wiley & Sons, 2001, p.306-318

Discussion

Course project introduction after 5-min break