Outline

- Non-parametric models for categorization: exemplars, neural networks
- Controlling complexity in statistical models

Bayesian classification

Image removed due to copyright considerations.

The task: Observe x generated from c_1 or c_2, compute:

$$p(c_1 \mid x) = \frac{p(x \mid c_1)\, p(c_1)}{p(x \mid c_1)\, p(c_1) + p(x \mid c_2)\, p(c_2)}$$

Different approaches vary in how they represent $p(x|c_j)$.

Non-parametric approaches

- Allow more complex form for $p(x|c_j)$, to be determined by the data.
- E.g., kernel density estimation:

$$p(x \mid c_j) \propto \frac{1}{n} \sum_{k=1}^{n} e^{-\|x - x^k\|^2 / (2\sigma^2)}$$

Image removed due to copyright considerations.

- Equivalent to exemplar model
  - Observe n examples: $x^1, \ldots, x^k$
  - Smoothness ("specificity"): σ

$$p(x \mid c_j) \propto \frac{1}{n}\sum_{k=1}^{n} e^{-\|x-x^k\|^2/(2\sigma^2)}$$

- Equivalent to exemplar model
  - Observe n examples: $x^1, \ldots, x^k$
  - Smoothness ("specificity"): σ

Image removed due to copyright considerations.

- Learning: Bayesian framework
  - Hypothesis space is all smooth density functions f.
  - Maximize $p(f \mid x^1, \ldots, x^k)$.
    - Prior $p(f)$ favors larger σ.
    - Likelihood $p(x^1, \ldots, x^k \mid f)$ favors smaller σ.

Nearest neighbor classification

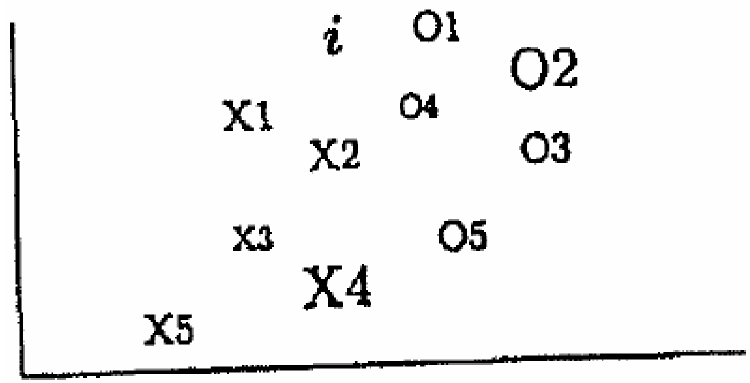
Image removed due to copyright considerations.

Theorem: In the limit of infinite data, classification according to nearest neighbor is approximately Bayes-optimal.

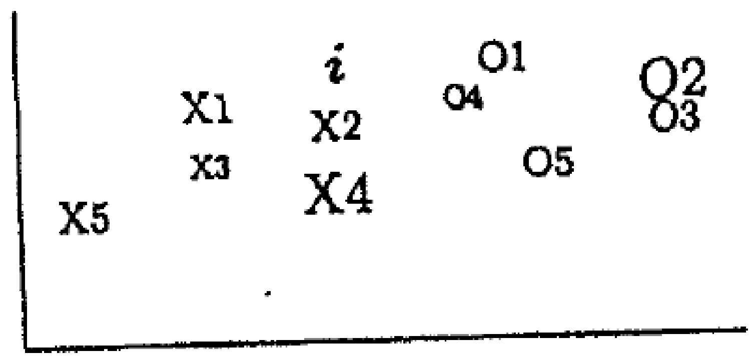
Nosofsky's exemplar model

- Motivating example:

Before learning



After learning:
selective attention



Nosofsky's exemplar model

- Math:
  - Probability of responding category J to stimulus i: Image removed due to copyright considerations.
  - Similarity of stimulus i to exemplar j:

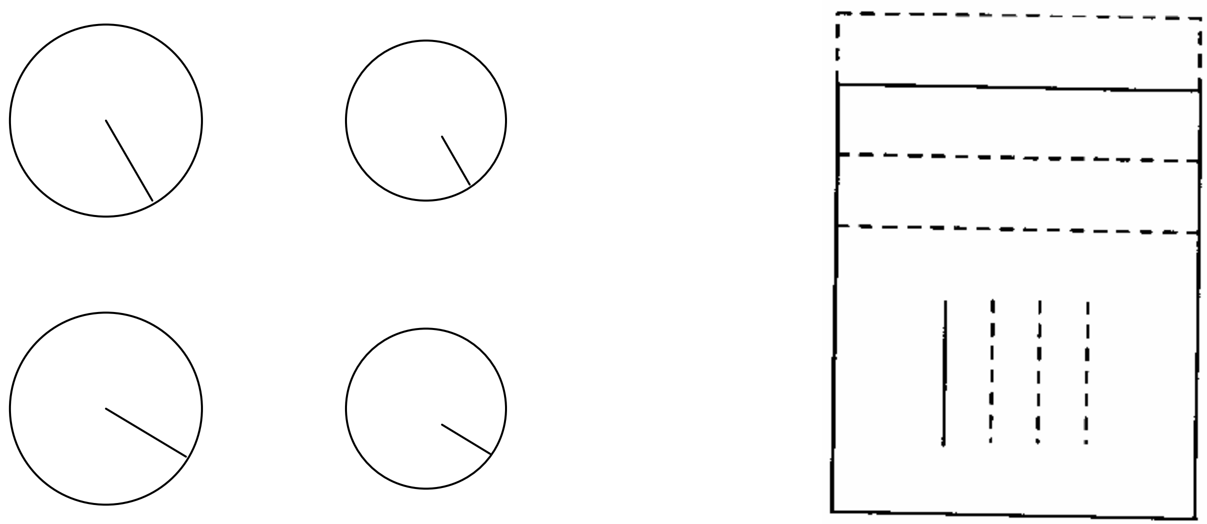
    Image removed due to copyright considerations.
  - Distance from stimulus i to exemplar j:

    Image removed due to copyright considerations.

w_m: attentional weight for dimension m ~ inverse variance

Concept learning experiments

- Simple artificial stimuli, e.g.



- Learn to discriminate members of two mutually exclusive categories, with repeated presentations of a small set of stimuli.

Six types of classifications:

Image removed due to copyright considerations.

Accurate fits to human data

Image removed due to copyright considerations.

How are attentional weights determined?

- By the modeler: tune to fit behavioral data

How are attentional weights determined?

- By the modeler: tune to fit behavioral data
- By the learner: *discriminative* learning
  - Maximize discriminability
    of the training set.

Image removed due to copyright considerations.

Generative vs. Discriminative Models

- Generative approach:
  - Separately model class-conditional densities $p(x \mid c_j)$ and priors $p(c_j)$.
  - Use Bayes' rule to compute posterior probabilities:

$$p(c_1 \mid x) = \frac{p(x \mid c_1)\, p(c_1)}{p(x \mid c_1)\, p(c_1) + p(x \mid c_2)\, p(c_2)}$$

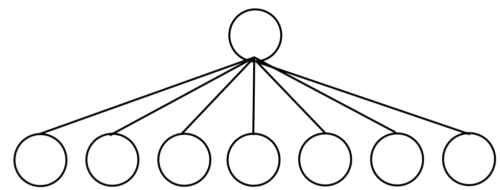
- Discriminative approach:
  - Directly model posterior probabilities $p(c_j \mid x)$

Generative vs. Discriminative

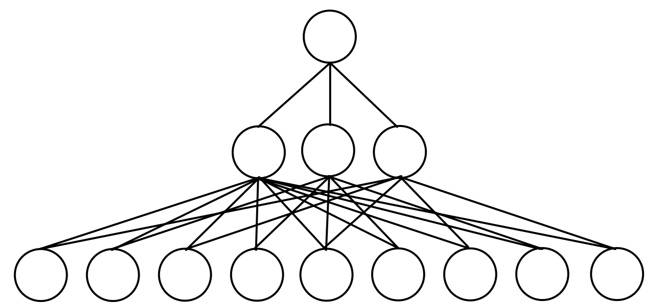
Image removed due to copyright considerations.

Discriminative methods based on function approximation

- Perceptrons

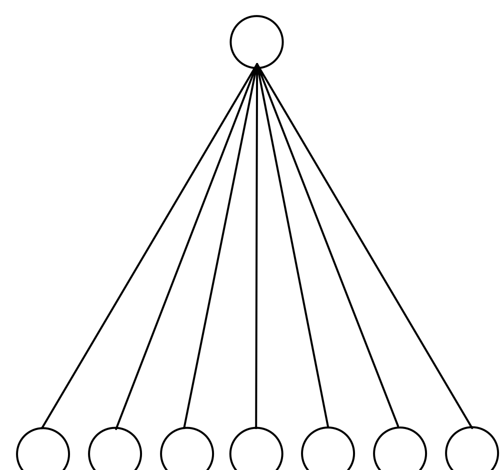


- Neural networks



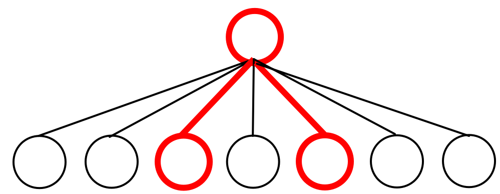
- Support vector machines

Image removed due to copyright considerations.



Discriminative methods based on function approximation

- Perceptrons



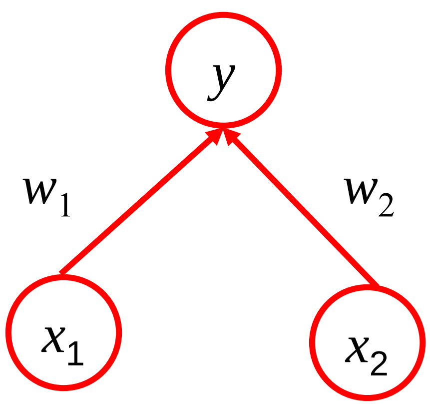


Image removed due to copyright considerations.

$$y = \theta(w_1 x_1 + w_2 x_2)$$

$$\theta(z) = 1/(1+\exp(-z))$$

Weight space

Image removed due to copyright considerations.

The perceptron hypothesis space

- Linearly separable classes

Image removed due to copyright considerations.

Perceptron learning

- Gradient descent on error surface in weight space:

Image removed due to copyright considerations.

Perceptron learning

- Gradient descent on error surface in weight space:

$$Error = y^* - y$$

$$y^* = \text{correct output}$$

$$E = \frac{1}{2} Error^2$$

$$y = \theta(\sum_j w_j x_j)$$

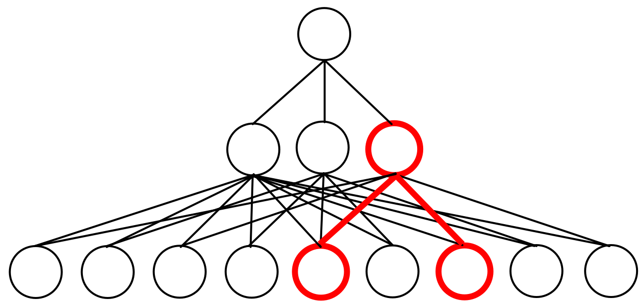
$$w_j \leftarrow w_j - \alpha \times \frac{\partial E}{\partial w_j}$$

$$\frac{\partial E}{\partial w_j} = Error \times \frac{\partial Error}{\partial w_j} \quad = -Error \times \frac{\partial y}{\partial w_j}$$

$$= -Error \times \theta'(\sum_j w_j x_j) \times x_j = -(y^* - y) \times x_j \times \theta'(\sum_j w_j x_j)$$

Discriminative methods based on function approximation

- Neural networks



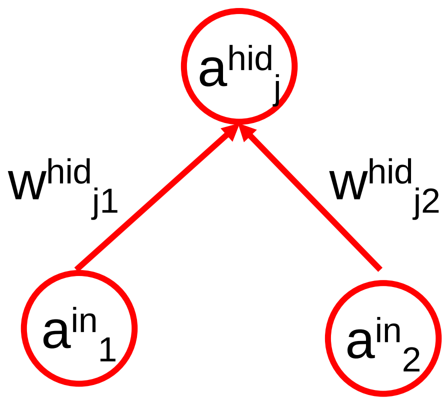


Image removed due to copyright considerations.

The benefit of hidden units

ridge = θ(sigmoid+sigmoid)   bump = θ(ridge+ridge)

Image removed due to copyright considerations.

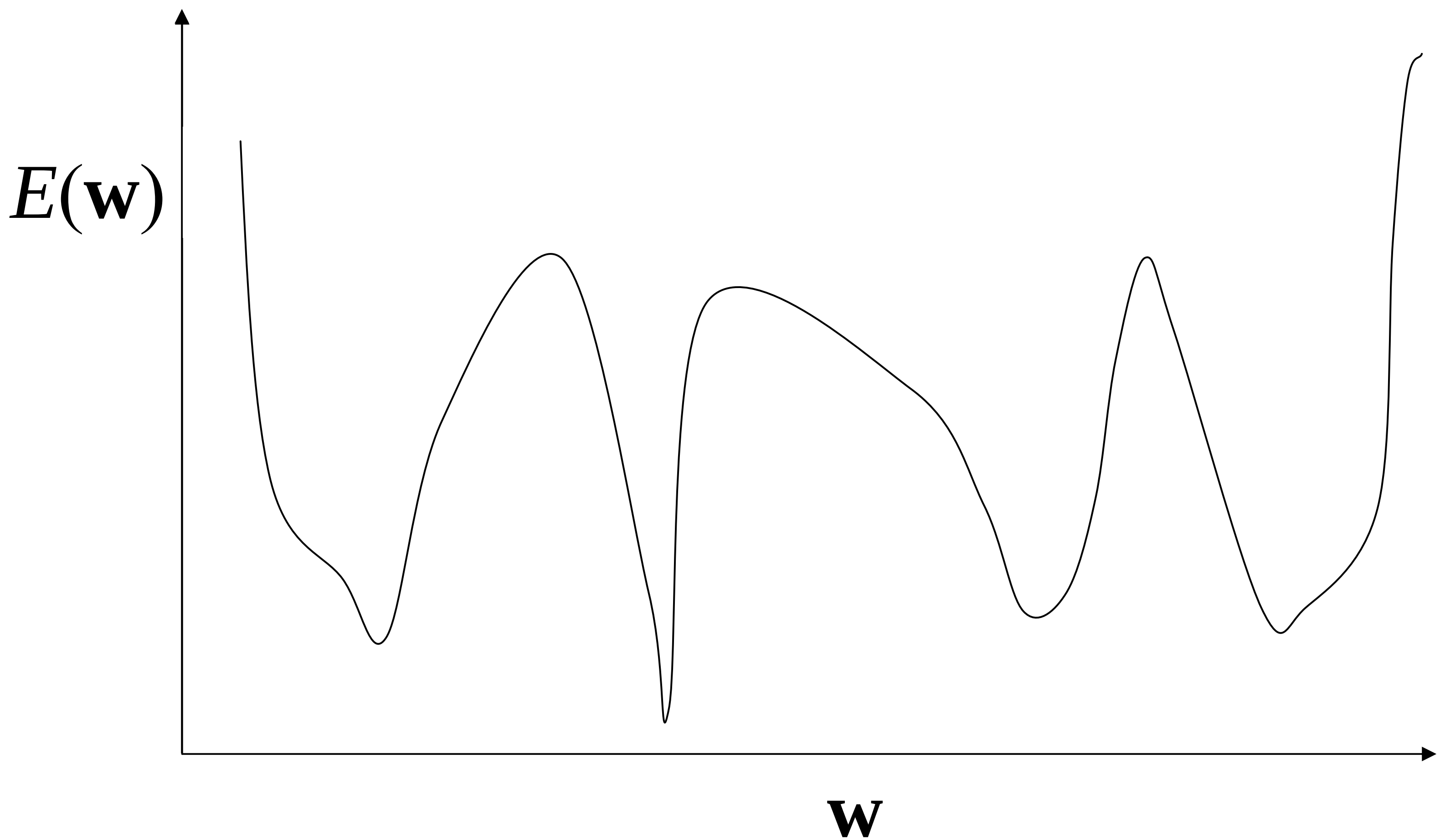
Neural network learning

- Gradient descent on error surface in weight space ("backpropagation"):

Image removed due to copyright considerations.

Neural network learning

- Gradient descent on error surface in weight space ("backpropagation"):



Backpropagation as a model of human learning?

- Kruschke: Are neural networks trained with backpropagation a good model of human category learning?
- Originally, backpropagation was not intended as a precise model of learning.
  - Rather, a tool for learning representations.
  - But does it learn the right kind of representations?

Two learning tasks

"Filtration": easy to learn

Image removed due to copyright considerations.

"Condensation": hard to learn

Image removed due to copyright considerations.

Human learning data

Image removed due to copyright considerations.

Conventional neural network

Image removed due to copyright considerations.

Model versus Data

Image removed due to copyright considerations.

People

Backprop

ALCOVE network

Image removed due to copyright considerations.

Differences between the models

- Dimension-specific attentional weights
- Hidden unit activation functions

Image removed due to copyright considerations.

Model versus Data

Image removed due to copyright considerations.

People

ALCOVE

Model versus Data

Image removed due to copyright considerations.

People

Backprop + attentional weights (c.f. ARD)

“Catastrophic forgetting”

- Stimuli for category-learning experiment

Image removed due to copyright considerations.

Human learning data

Image removed due to copyright considerations.

ALCOVE

Image removed due to copyright considerations.

Conventional neural network

Image removed due to copyright considerations.

Questions about neural networks

- Why do they have such a bad rap?
- To what extent are neural networks brain-like?
- They take a long time to train. Is that a good thing or a bad thing from the standpoint of cognitive modeling?