PSYCO 452

Week 7: The Analog Perceptron

- Intuitive Statistics
- Digital vs Analog Perceptrons
- Bayesian Probability
- Bayes' Theorem and Networks

Course Trajectory

When	What
~~Weeks 1-3~~	~~Basics of three architectures (DAM, perceptron, MLP)~~
~~Weeks 4-6~~	~~Cognitive science of DAMs and perceptrons~~
Week 7	**Connectionism and Cognitive Psychology**
Weeks 8-10	Interpreting MLPs
Weeks 11-13	Case studies (interpretations, applications, architectures)

Laplace's Demon

- **To an agent with knowledge of all causal relationships "nothing would be uncertain and the future, as the past, would be present to its eyes" (Laplace, 1814).**
- **Imperfect, humans must accept and adapt to uncertainty**
- **Probability theory is a means for doing this**
- **"The theory of probabilities is at bottom nothing but common sense reduced to calculus; it enables us to appreciate with exactness that which accurate minds feel with a sort of instinct for which oftimes they are unable to account."**

The Intuitive Statistician

- **20th century research uses probability and statistics to define norms to which human reasoning can be compared**
- **"[Our] psychological research consists of examining the relation between inferences made by man and corresponding optimal inferences as would be made by 'statistical man'" (Peterson & Beach, 1967, p. 29)**

Lee Roy Beach

Probability Theory Is Key

- **Egon Brunswik used probabilistic notions as the keys to his lens theory**

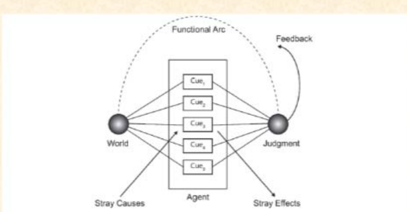
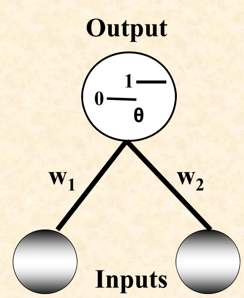


Figure 1-1. Brunswik's lens model of the mapping between the world and actions upon it. This figure is a variation of Brunswik (1952) Figure 1.

Egon Brunswik

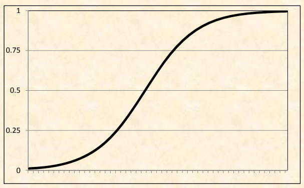
Digital Perceptron

- **The traditional perceptron has digital output and is equivalent to a trainable McCulloch-Pitts neuron**
- **Activation function is a step function or Heaviside equation with threshold θ**



The Logistic Equation

- But modern perceptrons typically use an analog activation function, the logistic
- Typically connectionists ignore this, and treat it as digital, by training outputs to the extremes of the logisitc



$$a_i = \frac{1}{1+e^{(-(net_i+\theta))}}$$

Analog Behavior

- But the analog nature of the activation function can be valuable
- Activity can match the probability of reinforcement, as shown by Dawson et al. (2009)
- What can be gained by exploring the analog properties of a modern perceptron?

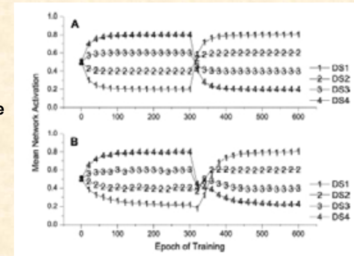


Fig. 1. Average responses of ten different perceptrons to each of the four stimuli as a function of training epoch. (a) Responses for networks from the first simulation which used standard training procedures. (b) Responses for networks from the second simulation which used an operant training procedure.

Perceptron As Lens

- We can use the analog perceptron as a candidate architecture for Brunswik's probabilistic lens model

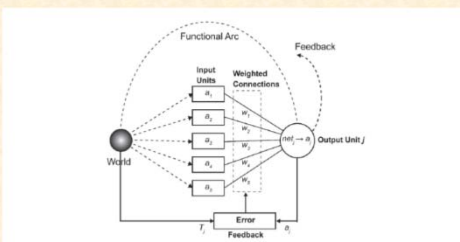


Figure 1-6. The basic components of a perceptron, placed in the context of Brunswik's lens model from Figure 1-1. See text for details.

a priori Probability

- One class of probability problems are *a priori* problems
- Knowing the cause, you make an inference about future events
- Example:
  - Cause:
    - box of marbles, 25% white, 75% black, sampled one at a time with replacement
  - Inference:
    - What is likelihood of drawing WWWBB when 5 selections are made from the box?

a posteriori Probability

- The inverse probability problem is to reason backwards from event to its (unknown) cause
- Example:
  - Evidence:
    - Draw WWWBB marbles
  - Hypothesis about cause:
    - What is likelihood draw came from bag of marbles, 25% white, 75% black?
- That is, what is P(H|E)?
- In general, one must consider the different *a posteriori* probabilities of a number of different competing hypothesis
- Simplest case – 2 hypotheses

Bayes' Rule

- In the simplest case, consider probabilities involving two hypothetical causes, H and ~H
- The posterior probability P(H|E) is given by a simple ratio:

$$P(H|E) = \frac{P(E|H)\cdot P(H)}{\big(P(E|H)\cdot P(H)\big) + \big(P(E|\sim H)\cdot P(\sim H)\big)}$$

$$= \frac{1}{1+\frac{P(E|\sim H)\cdot P(\sim H)}{P(E|H)\cdot P(H)}}$$

Thomas Bayes

Case Study: Contingency

- According to contingency theory, learning occurs when stimulus provides information about the likelihood of a certain event occurring
- Simple contiguity is not enough
- "The notion of contingency differs from that of *pairings* in that the former includes not only what *is* paired with the CS but also what *is not* paired with the CS" (Rescorla, 1967, p. 76).


Robert Rescorla

Measuring Contingency

- In the simplest scenario, summarized by a 2X2 contingency table, contingency between variables is defined by ΔP
- ΔP = P(H|E) – P(H|~E) = a/(a+b) – c/(c+d) = (ad – bc)/((a + b)(c + d))
- Perceptrons compute ΔP if activity is analog (and logistic)!

	H	*~H*
E	*a*	*b*
~E	*c*	*d*


Lorraine Allan

Bayes' Rule In Action

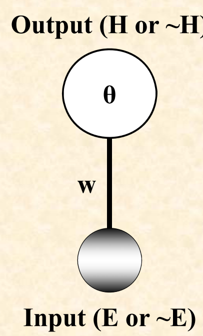
- What is probability that a patient has breast cancer given that their mammogram was positive?
- Note that this problem is based on a contingency table!
- $P(H|E) = \frac{1}{1+\frac{P(E|\sim H)\cdot P(\sim H)}{P(E|H)\cdot P(H)}} = \frac{1}{1+\frac{\left(\frac{b}{b+d}\right)\cdot\left(\frac{b+d}{a+b+c+d}\right)}{\left(\frac{a}{a+c}\right)\cdot\left(\frac{a+c}{a+b+c+d}\right)}} = \frac{1}{1+\left(\frac{b}{a}\right)}$

P(H|E) = 0.0776699029126214

	H	*~H*
E	*a = 8*	*b = 95*
~E	*c = 2*	*d = 895*

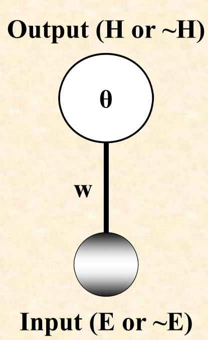
The Posterior Perceptron

- Train the simplest perceptron on the 1000 patterns defined by the cancer contingency table
- At the end of training:
  - *w* = 3.4656832744007
  - θ = -5.94807103049533
  - Input on: 0.0771021239932535
  - Input off: 0.00260407305517532
  - Difference between these two activities is ΔP
- Posterior perceptron learns to approximately agree with Bayes' rule!



Is The Perceptron Bayesian?

- Empirically, the simplest perceptron learns to behave like Bayes' rule
- Are the two formally equivalent?
- This issue can be explored using the contingency table to act as a link between the logistic and Bayes' rule



Equating Output With Input On

- Take the empirical link between perceptron behavior and Bayes' rule, and express it formally for when E is true:

$$P(H|E) = \frac{1}{1+e^{-(w+\theta)}} = \frac{1}{1+\left(\frac{b}{a}\right)} = \frac{1}{1+\frac{P(E|\sim H)\cdot P(\sim H)}{P(E|H)\cdot P(H)}} = \frac{1}{1+\frac{\left(\frac{b}{b+d}\right)\cdot\left(\frac{b+d}{a+b+c+d}\right)}{\left(\frac{a}{a+c}\right)\cdot\left(\frac{a+c}{a+b+c+d}\right)}}$$

$$e^{-(w+\theta)} = \frac{b}{a}$$

$$w + \theta = \ln(a) - \ln(b)$$

Equating Output With Input Off

- Take the empirical link between perceptron behavior and Bayes' rule, and express it formally for when E is false:

$$P(H|E) = \frac{1}{1+e^{-(\theta)}} = \frac{1}{1+\left(\frac{d}{c}\right)} = \frac{1}{1+\frac{P(\sim E|\sim H)\cdot P(\sim H)}{P(\sim E|H)\cdot P(H)}} = \frac{1}{1+\frac{\left(\frac{d}{b+d}\right)\cdot\left(\frac{b+d}{a+b+c+d}\right)}{\left(\frac{c}{a+c}\right)\cdot\left(\frac{a+c}{a+b+c+d}\right)}}$$

$$e^{-(\theta)} = \frac{d}{c}$$

$$\theta = \ln(c) - \ln(d)$$

Finding Weight

- The previous equations related network bias to Bayes' rule. We can use it to solve for the connection weight value too:

$$w = \ln(a) - \ln(b) - \theta$$

$$= \ln(a) - \ln(b) - \ln(c) + \ln(d)$$
$$= \ln(ad) - \ln(bc)$$
$$= \ln\left(\frac{ad}{bc}\right)$$

Bayesian Perceptron

- We can translate Bayes' rule into the weight and bias of the posterior perceptron
- Formal equivalence!
- Ideal
  - ***w*** **= 3.6292411877942051550032412410499**
  - **θ = -6.1036765377149100613613316417894**
  - ***w*** **+ θ = -2.4744353499207049063580904007395**
- Observed Example
  - ***w*** **= 3.4656832744007**
  - **θ = -5.94807103049533**
  - ***w*** **+ θ = -2.48238775609463**

Father of Bayesian Statistics

Extending Bayes' Theorem

- Bayes' theorem can be extended to cases in which more than one source of evidence is being used to signal probability

Table 4-3. General form of a 2X2X2 contingency table for two signals (X, Y) that can lead to a reward (R). Each lowercase letter in a cell stands for a frequency. For instance, a is the number of times that there is a reward when X and Y are both true, while e is the number of times that there is no reward when X and Y are both true.

	R			~R		
	Y	~Y	Sum	Y	~Y	Sum
X	a	b	a+b	e	f	e+f
~X	c	d	c+d	g	h	g+h
Sum	a+c	b+d	a+b+c+d	e+g	f+h	e+f+g+h

$$P(H|X \cap Y) = \frac{P(X \cap Y|H)\cdot P(H)}{\big(P(X \cap Y|H)\cdot P(H)\big) + \big(P(X \cap Y|\sim H)\cdot P(\sim H)\big)} \qquad (4\text{-}4)$$

Naïve Bayes

- The math of the extended Bayes' theorem can be simplified by making it 'blind' to interactions between variables
- This equation is called naïve Bayes
- Compare this to Equation 4-4

$$P(H|X \cap Y) = \frac{P(X|H)\cdot P(Y|H)\cdot P(H)}{\big(P(X|H)\cdot P(Y|H)\cdot P(H)\big) + \big(P(X|\sim H)\cdot P(Y|\sim H)\cdot P(\sim H)\big)} \qquad (4\text{-}5)$$

Perceptron Interpretation

- Formal analyses provide similar interpretations to the two-variable Bayesian perceptron

$$\begin{aligned} w_x &= -ln\left(\frac{(e+f)}{(a+b)}\right) - ln\left(\frac{(f+h)}{(b+d)}\right) - ln\left(\frac{(a+b+c+d)}{(e+f+g+h)}\right) - \theta \\ &= -ln\left(\frac{(e+f)}{(a+b)}\right) - ln\left(\frac{(f+h)}{(b+d)}\right) - ln\left(\frac{(a+b+c+d)}{(e+f+g+h)}\right) \\ &\quad - \left(-ln\left(\frac{(g+h)}{(c+d)}\right) - ln\left(\frac{(f+h)}{(b+d)}\right) - ln\left(\frac{(a+b+c+d)}{(e+f+g+h)}\right)\right) \\ &= ln\left(\frac{(g+h)}{(c+d)}\right) - ln\left(\frac{(e+f)}{(a+b)}\right) = ln\left(\frac{\frac{(g+h)}{(c+d)}}{\frac{(e+f)}{(a+b)}}\right) \\ &= ln\left(\frac{(a+b)\cdot(g+h)}{(e+f)\cdot(c+d)}\right) \end{aligned} \qquad (4\text{-}17)$$

$$\begin{aligned} w_y &= -ln\left(\frac{(g+h)}{(c+d)}\right) - ln\left(\frac{(e+g)}{(a+c)}\right) - ln\left(\frac{(a+b+c+d)}{(e+f+g+h)}\right) - \theta \\ &= -ln\left(\frac{(g+h)}{(c+d)}\right) - ln\left(\frac{(e+g)}{(a+c)}\right) - ln\left(\frac{(a+b+c+d)}{(e+f+g+h)}\right) \\ &\quad - \left(-ln\left(\frac{(g+h)}{(c+d)}\right) - ln\left(\frac{(f+h)}{(b+d)}\right) - ln\left(\frac{(a+b+c+d)}{(e+f+g+h)}\right)\right) \\ &= ln\left(\frac{(f+h)}{(b+d)}\right) - ln\left(\frac{(e+g)}{(a+c)}\right) = ln\left(\frac{\frac{(f+h)}{(b+d)}}{\frac{(e+g)}{(a+c)}}\right) \\ &= ln\left(\frac{(a+c)\cdot(f+h)}{(b+d)\cdot(e+g)}\right) \end{aligned} \qquad (4\text{-}18)$$

$$\begin{aligned} \theta &= -ln\left(\frac{(g+h)}{(c+d)}\right) - ln\left(\frac{(f+h)}{(b+d)}\right) - ln\left(\frac{(a+b+c+d)}{(e+f+g+h)}\right) \\ &= ln\left(\frac{(c+d)}{(g+h)}\right) + ln\left(\frac{(b+d)}{(f+h)}\right) - ln\left(\frac{(a+b+c+d)}{(e+f+g+h)}\right) \end{aligned} \qquad (4\text{-}19)$$

Three Cue Test Case

- Train perceptrons where probability of reward is signaled by three cues
- In some conditions let two cues interact (AND, XOR)
- Also manipulate reward probability associated with the interaction

Figure 7-1. A view of the card tableau that participants in the card-choice task interact with. In this particular image a participant has already made 16 different card choices, and has been rewarded for 6 of them.

Perceptron Performance

- Perceptrons, not surprisingly, have trouble with interactions, particularly when reward probability makes conditional dependence very hard

Table 5-18. The mean probability estimation performance (with standard deviations in parentheses) of perceptrons as a function of problem type and level of reward. Probability estimation performance is operationalized as the squared correlation between the perceptrons responses for each of eight different stimuli and the actual reward probabilities for these same stimuli. Each mean is based upon the performance of 100 different perceptrons.

	High Reward	Low Reward
AND of Cues B and C	0.82 (0.03)	0.95 (0.01)
XOR of Cues B and C	0.52 (0.05)	0.90 (0.03)

Table 5-20. The mean probability estimation performance (with standard deviations) of perceptrons in each of the five simulations that are described in Chapter 5. Probability estimation performance is operationalized as the squared correlation between the response generated by a perceptron to each of eight possible stimuli and the actual probability of reward associated with each. Each mean is based upon the performance of 100 different perceptrons.

	Independent Cues	High Reward AND	High Reward XOR	Low Reward AND	Low Reward XOR
Mean R^2	0.93	0.82	0.52	0.95	0.90
SD	0.04	0.03	0.05	0.01	0.03

Human Performance

- Human performance on an analogous task is very similar, suggesting that during probability learning people are like naïve Bayesians!

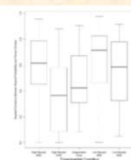


Figure 7-3. Box plots showing the distribution of R^2 values between the choice behavior of individuals and ideal reward probabilities for each of the conditions in the card-choice task. Each box plot summarizes the responses of 40 different participants.

Table 7-4. The mean squared correlation between individual participant's choices of stimuli and the ideal reward probabilities. The SE row provides the standard errors of these means, each of which is based upon 40 different participants.

	High Reward AND	High Reward XOR	Independent Cues	Low Reward AND	Low Reward XOR
Mean R^2	0.569	0.380	0.480	0.626	0.550
SE	0.043	0.045	0.037	0.043	0.040