COMP9444
Neural Networks and Deep Learning

4. Variations on Backprop

Textbook, Sections 3.1-3.6, 3.9-3.11, 5.2.2, 5.5, 8.3

Outline

- Probability (3.1-3.6, 3.9.3, 3.10)
- Cross Entropy (5.5)
- Bayes' Rule (3.11)
- Weight Decay (5.2.2)
- Momentum (8.3)

Probability (3.1)

Begin with a set Ω – the sample space (e.g. 6 possible rolls of a die)

$\omega \in \Omega$ is a sample point/possible world/atomic event

A probability space or probability model is a sample space
with an assignment $P(\omega)$ for every $\omega \in \Omega$ s.t.

$0 \leq P(\omega) \leq 1$

$\sum_{\omega} P(\omega) = 1$
e.g. $P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = \frac{1}{6}$.

An event A is any subset of Ω

$$P(A) = \sum_{\{\omega \in A\}} P(\omega)$$

e.g. $P(\text{die roll} < 4) = P(1) + P(2) + P(3) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}$

Random Variables (3.2)

A random variable (r.v.) is a function from sample points to some range (e.g. the Reals or Booleans)

For example, $\texttt{Odd}(3) = \texttt{true}$.

P induces a probability distribution for any r.v. X:

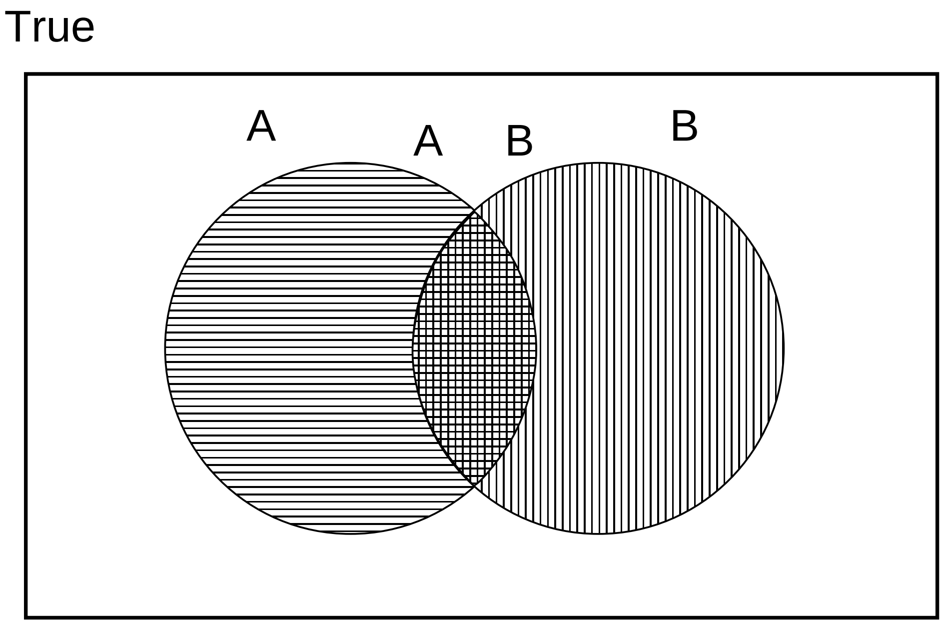
$$P(X = x_i) = \sum_{\{\omega : X(\omega) = x_i\}} P(\omega)$$

e.g., $P(\texttt{Odd} = \texttt{true}) = P(1) + P(3) + P(5) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}$

Probability and Logic

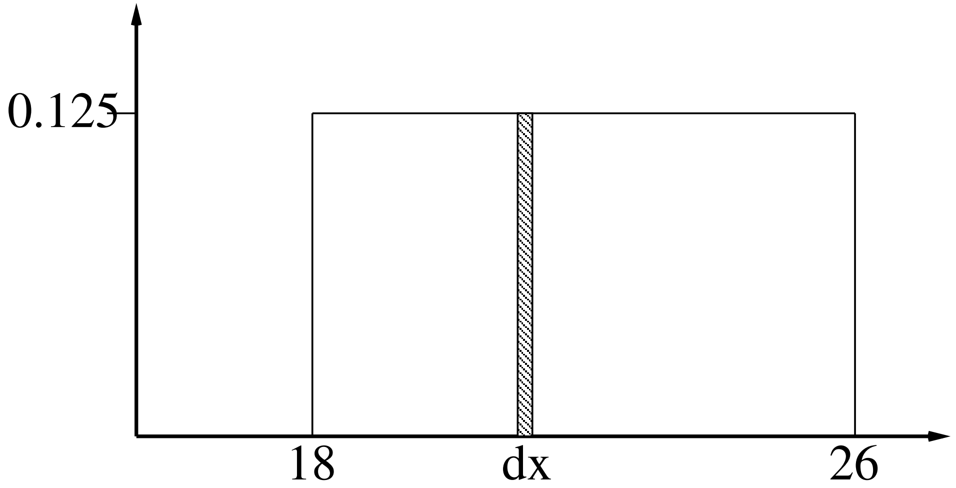
Logically related events must have related probabilities

For example, $P(a \vee b) = P(a) + P(b) - P(a \wedge b)$



Probability for Continuous Variables

e.g. $P(X = x) = U[18, 26](x)$ = uniform density between 18 and 26



Here P is a density; integrates to 1.

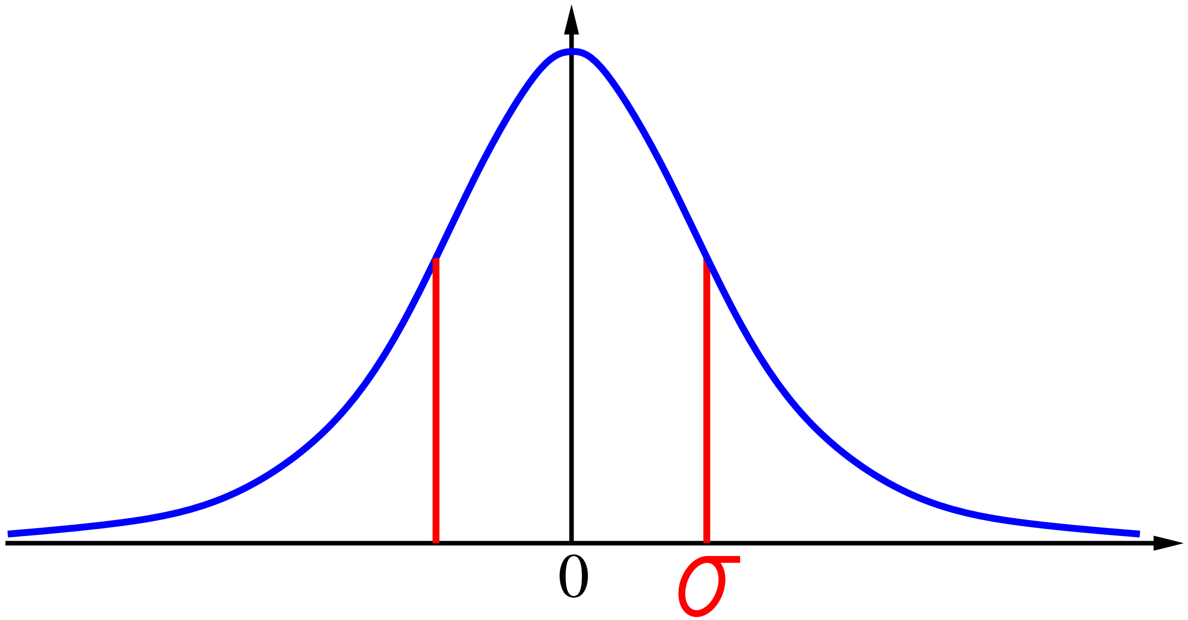
$P(X = 20.5) = 0.125$ really means

$$\lim_{dx \to 0} P(20.5 \leq X \leq 20.5 + dx)/dx = 0.125$$

Gaussian Distribution (3.9.3)

$\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$

$$P(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}$$



Variations on Backprop

- Cross Entropy $E = -t\log(z) - (1-t)\log(1-z)$
  - problem: least squares error function unsuitable for classification, where target = 0 or 1
  - mathematical theory: maximum likelihood
  - solution: replace with cross entropy error function
- Weight Decay
  - problem: weights "blow up", and inhibit further learning
  - mathematical theory: Bayes' rule
  - solution: add weight decay term to error function
- Momentum
  - problem: weights oscillate in a "rain gutter"
  - solution: weighted average of gradient over time

Cross Entropy

For classification tasks, target t is either 0 or 1, so better to use

$$\boxed{E = -t\log(z) - (1-t)\log(1-z)}$$

This can be justified mathematically, and works well in practice – especially when negative examples vastly outweigh positive ones.

It also makes the backprop computations simpler

$$\frac{\partial E}{\partial z} = -\frac{t}{z} + \frac{1-t}{1-z} = \frac{t-z}{z(1-z)}$$

$$\frac{\partial E}{\partial z} = \frac{z-t}{z(1-z)}$$

$$\text{if} \quad z = \frac{1}{1+e^{-s}},$$

$$\frac{\partial E}{\partial s} = \frac{\partial E}{\partial z}\frac{\partial z}{\partial s} = z - t$$

Maximum Likelihood (5.5)

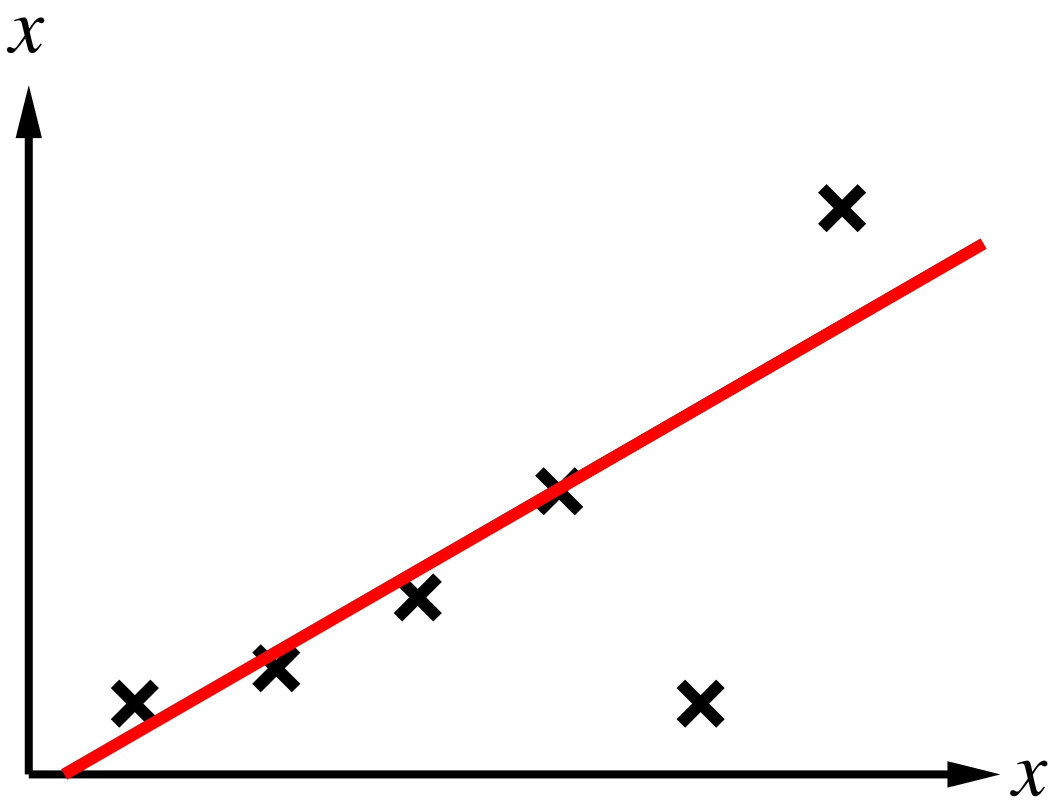
H is a class of hypotheses

$P(D|h)$ = probability of data D being generated under hypothesis $h \in H$.

$\log P(D|h)$ is called the likelihood.

ML Principle: Choose $h \in H$ which maximizes the likelihood,
i.e. maximizes $P(D|h)$ [or, maximizes $\log P(D|h)$]

Least Squares Line Fitting



Derivation of Least Squares

Suppose data generated by a linear function h, plus Gaussian noise with standard deviation σ.

$$
\begin{aligned}
P(D|h) &= \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{1}{2\sigma^2}(d_i - h(x_i))^2} \\
\log P(D|h) &= \sum_{i=1}^{m} -\frac{1}{2\sigma^2}(d_i - h(x_i))^2 - \log(\sigma) - \frac{1}{2}\log(2\pi) \\
h_{ML} &= \mathrm{argmax}_{h\in H}\, \log P(D|h) \\
&= \mathrm{argmin}_{h\in H} \sum_{i=1}^{m} (d_i - h(x_i))^2
\end{aligned}
$$

(Note: we do not need to know σ)

Derivation of Cross Entropy

For classification tasks, d is either 0 or 1.

Assume D generated by hypothesis h as follows:

$$
\begin{aligned}
P(1|h(x_i)) &= h(x_i) \\
P(0|h(x_i)) &= (1-h(x_i)) \\
\text{i.e.} \quad P(d_i|h(x_i)) &= h(x_i)^{d_i}(1-h(x_i))^{1-d_i}
\end{aligned}
$$

then

$$
\begin{aligned}
\log P(D|h) &= \sum_{i=1}^{m} d_i \log h(x_i) + (1-d_i)\log(1-h(x_i)) \\
h_{ML} &= \text{argmax}_{h\in H} \sum_{i=1}^{m} d_i \log h(x_i) + (1-d_i)\log(1-h(x_i))
\end{aligned}
$$

(Can be generalized to multiple classes.)

Joint Probability Distribution

We assume there is some underlying joint probability distribution over the three random variables Toothache, Cavity and Catch, which we can write in the form of a table:

	toothache		*¬ toothache*	
	catch	*¬ catch*	*catch*	*¬ catch*
ca it	**.1 8**	**. 12**	**. 2**	**. 8**
¬ ca it	**. 16**	**. 64**	**.144**	**.5 6**

Note that the sum of the entries in the table is 1.0.

For any proposition ϕ, sum the atomic events where it is true:

$$P(\phi) = \sum_{\omega:\omega\models\phi} P(\omega)$$

Inference by Enumeration

Start with the joint distribution:

	toothache		*¬ toothache*	
	catch	*¬ catch*	*catch*	*¬ catch*
ca it	**.1 8**	**. 12**	**. 2**	**. 8**
¬ ca it	**. 16**	**. 64**	**.144**	**.5 6**

For any proposition ϕ, sum the atomic events where it is true:

$P(\phi) = \sum_{\omega:\omega\models\phi} P(\omega)$

$P(toothache) = 0.108 + 0.012 + 0.016 + 0.064 = 0.2$

Inference by Enumeration

	toothache		*¬ toothache*	
	catch	*¬ catch*	*catch*	*¬ catch*
ca it	**.1 8**	**. 12**	**. 2**	**. 8**
¬ ca it	**. 16**	**. 64**	**.144**	**.5 6**

For any proposition ϕ, sum the atomic events where it is true:

$P(\phi) = \sum_{\omega:\omega\models\phi} P(\omega)$

$P(\texttt{cavity} \vee \texttt{toothache})$
$= 0.108 + 0.012 + 0.072 + 0.008 + 0.016 + 0.064 = 0.28$

Conditional Probability (3.5-3.6)

If we consider two random variables a and b, with $P(b) \neq 0$, then the conditional probability of a given b is

$$P(a|b) = \frac{P(a \wedge b)}{P(b)}$$

Alternative formulation: $P(a \wedge b) = P(a|b)P(b) = P(b|a)P(a)$

When we consider a sequence of random variables at successive time steps, they can be chained together using this formula repeatedly:

$$\begin{aligned} P(X_n, \dots, X_1) &= P(X_n \,|\, X_{n-1}, \dots, X_1) P(X_{n-1}, \dots, X_1) \\ &= P(X_n \,|\, X_{n-1}, \dots, X_1) P(X_{n-1} \,|\, X_{n-2}, \dots, X_1) \\ &= \dots = \prod_{i=1}^{n} P(X_i \,|\, X_{i-1}, \dots, X_1) \end{aligned}$$

Conditional Probability by Enumeration

	toothache		¬ *toothache*	
	catch	¬ *catch*	*catch*	¬ *catch*
ca it	**.1 8**	**. 12**	**. 2**	**. 8**
¬ *ca it*	**. 16**	**. 64**	**.144**	**.5 6**

$$P(\neg \texttt{cavity} \,|\, \texttt{toothache}) = \frac{P(\neg \texttt{cavity} \wedge \texttt{toothache})}{P(\texttt{toothache})}$$

$$= \frac{0.016 + 0.064}{0.108 + 0.012 + 0.016 + 0.064} = 0.4$$

Bayes' Rule (3.11)

The formula for conditional probability can be manipulated to find a relationship when the two variables are swapped:

$$P(a \wedge b) = P(a \,|\, b)P(b) = P(b \,|\, a)P(a)$$

$$\rightarrow \text{Bayes' rule } P(a \,|\, b) = \frac{P(b \,|\, a)P(a)}{P(b)}$$

This is often useful for assessing the probability of an underlying cause after an effect has been observed:

$$P(\text{Cause}|\text{Effect}) = \frac{P(\text{Effect}|\text{Cause})P(\text{Cause})}{P(\text{Effect})}$$

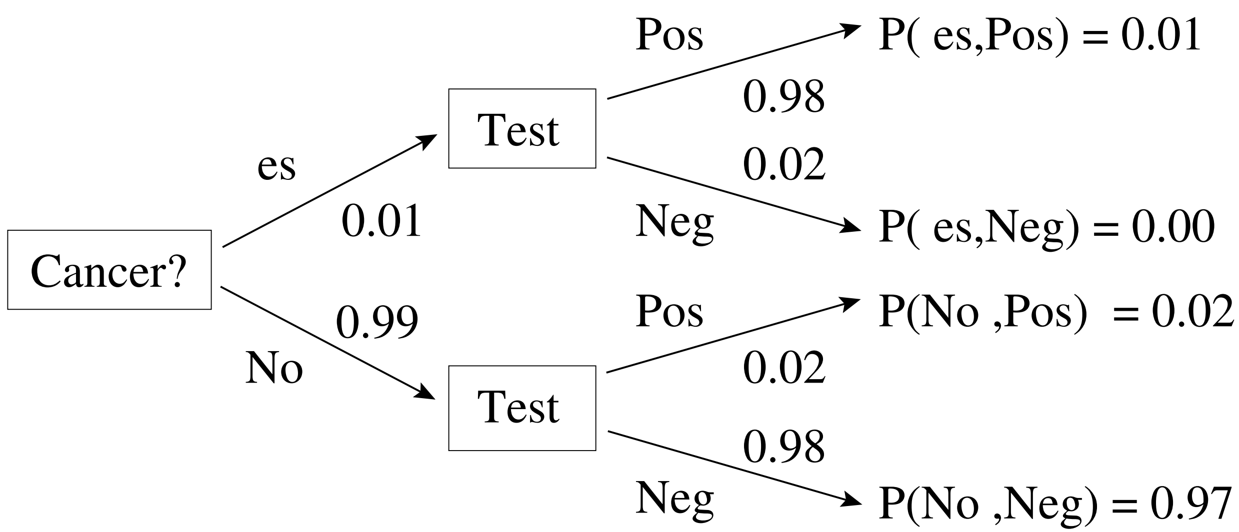
Example: Medical Diagnosis

Question: Suppose we have a 98% accurate test for a type of cancer which occurs in 1% of patients. If a patient tests positive, what is the probability that they have the cancer?

Answer: There are two random variables: Cancer (true or false) and Test (positive or negative). The probability is called a prior, because it represents our estimate of the probability before we have done the test (or made some other observation). We interpret the statement that the test is 98% accurate to mean:

$$P(\texttt{positive} \mid \texttt{cancer}) = 0.98, \quad \text{and} \quad P(\texttt{negative} \mid \neg\texttt{cancer}) = 0.98$$

Bayes' Rule



$$P(\texttt{cancer}\,|\,\texttt{positive}) = \frac{P(\texttt{positive}\,|\,\texttt{cancer})P(\texttt{cancer})}{P(\texttt{positive})}$$

$$= \frac{0.98*0.01}{0.98*0.01+0.2*0.99} = \frac{0.01}{0.01+0.02} = \frac{1}{3}$$

Bayes Rule in Machine Learning

H is a class of hypotheses

$P(D|h)=$ probability of data D being generated under hypothesis $h \in H$.

$P(h|D)=$ probability that h is correct, given that data D were observed.

Bayes' Theorem:

$$\begin{aligned} P(h|D)P(D) &= P(D|h)P(h) \\ P(h|D) &= \frac{P(D|h)P(h)}{P(D)} \end{aligned}$$

$P(h)$ is called the prior.

Weight Decay (5.2.2)

Assume that small weights are more likely to occur than large weights, i.e.

$$P(w) = \frac{1}{Z} e^{-\frac{\lambda}{2} \sum_j w_j^2}$$

where Z is a normalizing constant. Then the cost function becomes:

$$\boxed{E = \frac{1}{2} \sum_i (z_i - t_i)^2 + \frac{\lambda}{2} \sum_j w_j^2}$$

This can prevent the weights from "saturating" to very high values.

Problem: need to determine λ from experience, or empirically.

Momentum (8.3)

If landscape is shaped like a "rain gutter", weights will tend to oscillate without much improvement.

Solution: add a momentum factor

$$\boxed{\begin{aligned} \delta w &\leftarrow \alpha\,\delta w + (1-\alpha)\frac{\partial E}{\partial w} \\ w &\leftarrow w - \eta\,\delta w \end{aligned}}$$

Hopefully, this will dampen sideways oscillations but amplify downhill motion by $\frac{1}{1-\alpha}$.

Conjugate Gradients

Compute matrix of second derivatives $\frac{\partial^2 E}{\partial w_i \partial w_j}$ (called the Hessian).

Approximate the landscape with a quadratic function (paraboloid).

Jump to the minimum of this quadratic function.

Natural Gradients (Amari, 1995)

Use methods from information geometry to find a “natural” re-scaling of the partial derivatives.

