COMP9444: Neural Networks and Deep Learning

Week 2a. Probability

Alan Blair

School of Computer Science and Engineering

June 3, 2024

Outline

- Probability and Random Variables (3.1-3.2)
- Probability for Continuous Variables (3.3)
- Gaussian Distributions (3.9.3)
- Conditional Probability (3.5)
- Bayes' Rule (3.11)
- Entropy and KL-Divergence (3.13)
- Continuous Distributions
- Wasserstein Distance

Probability (3.1)

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

Each $\omega \in \Omega$ 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 $\omega \in \Omega$ such that

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

$$\sum_{\omega} P(\omega) = 1$$

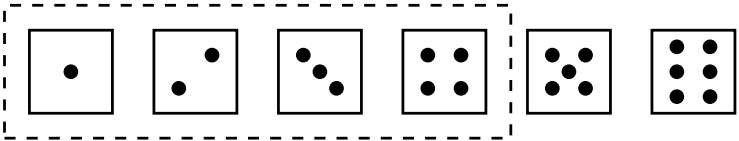
e.g. $P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = \frac{1}{6}$.

Random Events

A *random event* A is any subset of Ω

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

e.g. $P(\text{die roll} < 5) = P(1) + P(2) + P(3) + P(4) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{2}{3}$



Random Variables (3.2)

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

For example, `Odd`(3) = `true`

P induces a *probability distribution* for any random variable X:

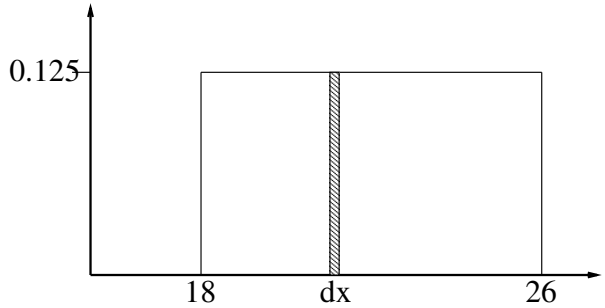
$$P(X = x_i) = \sum_{\{\omega : X(\omega) = x_i\}} P(\omega)$$

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 for Continuous Variables (3.3)

For continuous variables, P is a *density*; it integrates to 1.

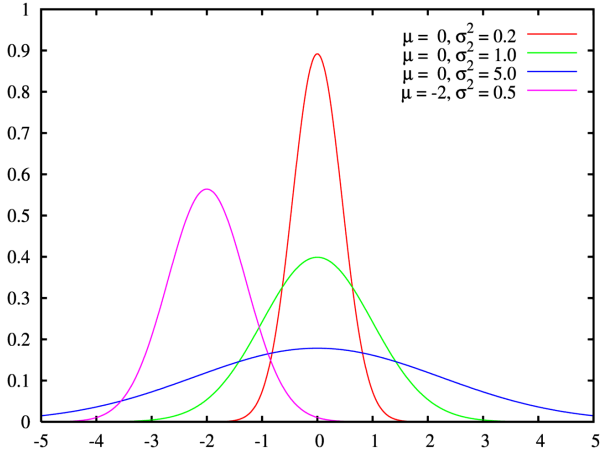
e.g. $P(X = x) = U[18, 26](x)$ = uniform density between 18 and 26



When we say $P(X = 20.5) = 0.125$, it really means

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

Gaussian Distribution (3.9.3)



$\mu =$ mean

$\sigma =$ standard deviation

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

Multivariate Gaussians

The d-dimensional multivariate Gaussian with mean μ and covariance Σ is given by

$$P_{\mu,\Sigma}(x) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} e^{-\frac{1}{2}(x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu)}$$

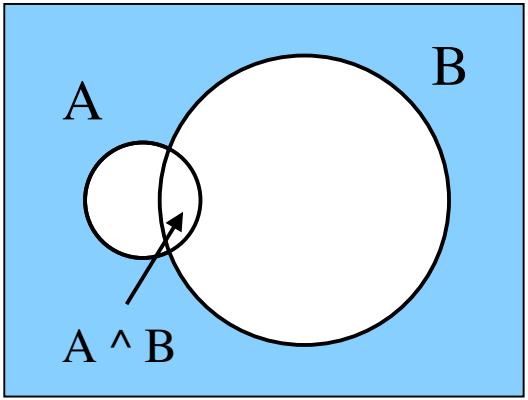
where $|\Sigma|$ denotes the determinant of Σ.

If $\Sigma = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2)$ is diagonal, the multivariate Gaussian reduces to

$$P_{\mu,\Sigma}(x) = \prod_i P_{\mu_i,\sigma_i}(x_i)$$

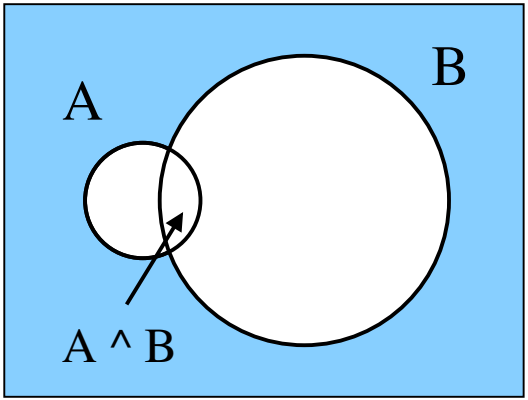
The Gaussian with $\mu = 0$, $\Sigma = \mathrm{I}$ is called the *Standard Normal* distribution.

Probability and Logic



Logically related events must have related probabilities
For example, $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$

Conditional Probability (3.5)



If $P(B) \neq 0$, then the *conditional probability* of A given B is

$$P(A \mid B) = \frac{P(A \wedge B)}{P(B)}$$

Bayes' Rule (3.11)

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

$$P(A \wedge B) = P(A \mid B)P(B) = P(B \mid A)P(A)$$

$$\rightarrow \textit{Bayes' rule} \quad P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

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

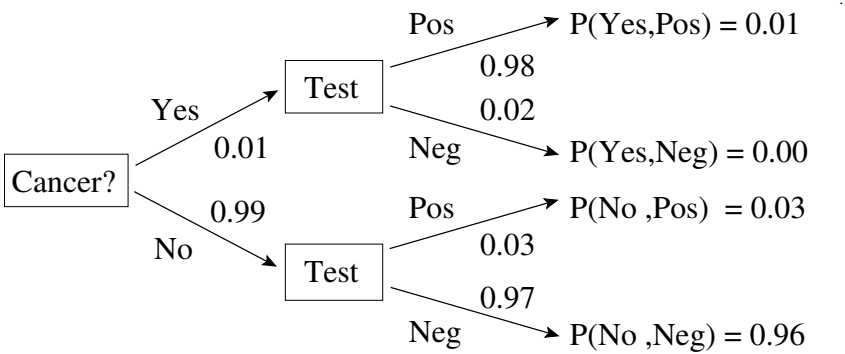
Question: Suppose we have a test for a type of cancer which occurs in 1% of patients. The test has a sensitivity of 98% and a specificity of 97%.
If a patient tests positive, what is the probability that they have the cancer?

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).

The *sensitivity* and *specificity* are interpreted as follows:

$$P(\texttt{positive} \mid \texttt{cancer}) = 0.98, \quad \text{and} \quad P(\texttt{negative} \mid \neg\texttt{cancer}) = 0.97$$

Bayes' Rule for Medical Diagnosis



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

$$= \frac{0.98 * 0.01}{0.98 * 0.01 + 0.03 * 0.99} = \frac{0.01}{0.01 + 0.03} = \frac{1}{4}$$

Example: Light Bulb Defects

Question: You work for a lighting company which manufactures 60% of its light bulbs in Factory A and 40% in Factory B. One percent of the light bulbs from Factory A are defective, while two percent of those from Factory B are defective. If a random light bulb turns out to be defective, what is the probability that it was manufactured in Factory A?

Answer: There are two random variables: Factory (A or B) and Defect (Yes or No).

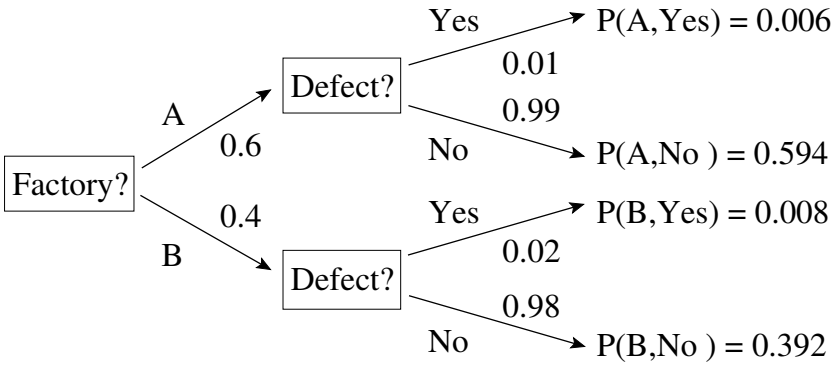
In this case, the prior is:

$$P(A) = 0.6, \qquad P(B) = 0.4$$

The conditional probabilities are:

$$P(\texttt{defect} \mid \texttt{A}) = 0.01, \quad \text{and} \quad P(\texttt{defect} \mid \texttt{B}) = 0.02$$

Bayes' Rule for Light Bulb Defects



$$P(\texttt{A} \mid \texttt{defect}) = \frac{P(\texttt{defect} \mid \texttt{A})P(\texttt{A})}{P(\texttt{defect})}$$

$$= \frac{0.01 * 0.6}{0.01 * 0.6 + 0.02 * 0.4} = \frac{0.006}{0.006 + 0.008} = \frac{3}{7}$$

Entropy and KL-Divergence (3.13)

The *entropy* of a discrete probability distribution $p = \langle p_1, \ldots, p_n \rangle$ is

$$H(p) = \sum_{i=1}^{n} p_i \left(-\log_2 p_i\right)$$

Given two probability distributions $p = \langle p_1, \ldots, p_n \rangle$ and $q = \langle q_1, \ldots, q_n \rangle$ on the same set Ω, the *Kullback-Leibler Divergence* between p and q is

$$D_{\mathrm{KL}}(p \,\|\, q) = \sum_{i=1}^{n} p_i \left(\log_2 p_i - \log_2 q_i\right)$$

KL-Divergence is like a "distance" from one probability distribution to another. But, it is not symmetric.

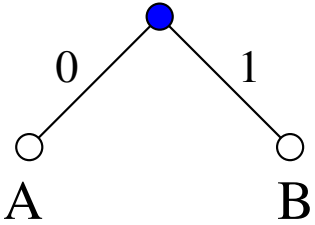
$$D_{\mathrm{KL}}(p \,\|\, q) \;\neq\; D_{\mathrm{KL}}(q \,\|\, p)$$

Entropy and Huffmann Coding

Entropy is the number of bits per symbol achieved by a (block) Huffman Coding scheme.

Example 1: $H(\langle 0.5, 0.5 \rangle) = 1$ bit.

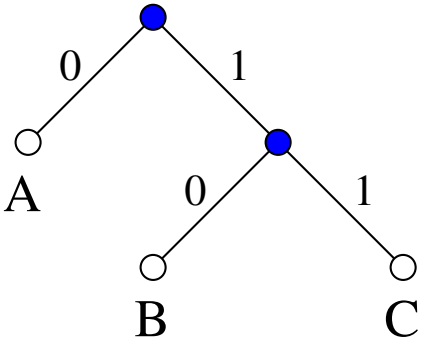
Suppose we want to encode, in zeros and ones, a long message composed of the letters `A` and `B`, which occur with equal frequency. This can be done efficiently by assigning `A=0`, `B=1`. In other words, one bit is needed to encode each letter.



Entropy and Huffmann Coding

Example 2: $H(\langle 0.5, 0.25, 0.25 \rangle) = 1.5$ bits.

Suppose we need to encode a message consisting of the letters `A`, `B` and `C`, where `B` and `C` occur equally often but `A` occurs twice as often as the other two letters.
In this case, an optimally efficient code would be `A=0`, `B=10`, `C=11`.
The average number of bits needed to encode each letter is 1.5.



Entropy and KL-Divergence

If the samples occur in some other proportion, we would need to "block" them together in order to encode them efficiently. But, the average number of bits required by the most efficient coding scheme is given by

$$H(\langle p_1, \ldots, p_n \rangle) = \sum_{i=1}^{n} p_i \left(-\log_2 p_i\right)$$

$\mathrm{D}_{\mathrm{KL}}(q \parallel p)$ is the number of *extra* bits we need to trasmit if we designed a code for $q()$ but it turned out that the samples were drawn from $p()$ instead.

$$D_{\mathrm{KL}}(p \parallel q) = \sum_{i=1}^{n} p_i \left(\log_2 p_i - \log_2 q_i\right)$$

Continuous Entropy and KL-Divergence

- the *entropy* of a continuous distribution $p()$ is

$$H(p) = \int_{\theta} p(\theta)(-\log p(\theta))\, d\theta$$

- *KL-Divergence* between two continuous distributions $p()$ and $q()$ is

$$\mathrm{D_{KL}}(p \parallel q) = \int_{\theta} p(\theta)(\log\ p(\theta) - \log\ q(\theta))\, d\theta$$

Entropy for Gaussian Distributions

Entropy of Gaussian with mean μ and standard deviation σ :

$$\frac{1}{2}(1 + \log(2\pi)) + \log(\sigma)$$

Entropy of a d-dimensional Gaussian $p()$ with mean μ and variance Σ :

$$H(p) = \frac{d}{2}(1 + \log(2\pi)) + \frac{1}{2}\log|\Sigma|$$

If $\Sigma = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2)$ is diagonal, the entropy is:

$$H(p) = \frac{d}{2}(1 + \log(2\pi)) + \sum_{i=1}^{d} \log(\sigma_i)$$

KL-Divergence for Gaussians

KL-Divergence between Gaussians $q()$, $p()$ with mean μ_1, μ_2 and variance Σ_1, Σ_2:

$$\mathrm{D}_{\mathrm{KL}}(q||p) = \frac{1}{2}\Big[(\mu_2 - \mu_1)^{\mathrm{T}}\Sigma_2^{-1}(\mu_2 - \mu_1) + \mathrm{Trace}(\Sigma_2^{-1}\Sigma_1) + \log\frac{|\Sigma_2|}{|\Sigma_1|} - d\Big]$$

In the case where $\mu_2 = 0$, $\Sigma_2 = \mathrm{I}$, the KL-Divergence simplifies to:

$$\mathrm{D}_{\mathrm{KL}}(q||p) = \frac{1}{2}\big[||\mu_1||^2 + \mathrm{Trace}(\Sigma_1) - \log|\Sigma_1| - d\big]$$

If $\Sigma_1 = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2)$ is diagonal, this reduces to:

$$\mathrm{D}_{\mathrm{KL}}(q||p) = \frac{1}{2}\big[||\mu_1||^2 + \sum_{i=1}^{d}(\sigma_i^2 - 2\log(\sigma_i) - 1)\big]$$

Wasserstein Distance

Another commonly used measure is the *Wasserstein Distance* which, for multivariate Gaussians, is given by

$$W_2(q,p)^2 = ||\mu_1 - \mu_2||^2 + \text{Trace}(\Sigma_1 + \Sigma_2 - 2(\Sigma_1\Sigma_2)^{\frac{1}{2}})$$

In the case where $\mu_2 = 0$, $\Sigma_2 = \text{I}$, the KL-Divergence simplifies to:

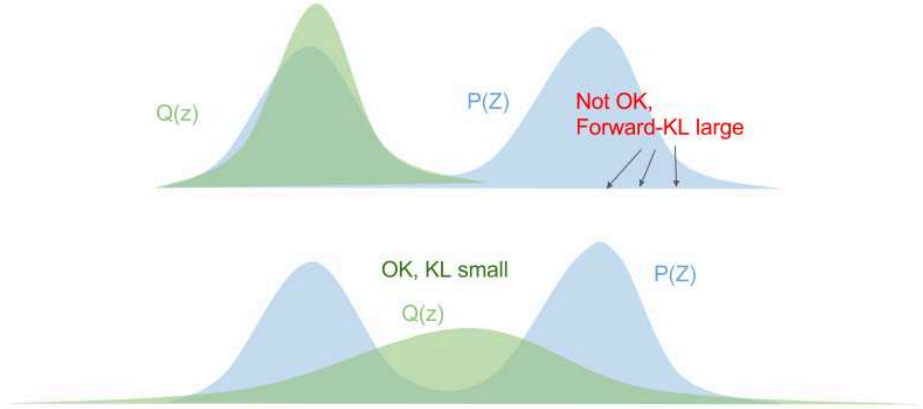
$$W_2(q,p)^2 = ||\mu_1||^2 + d + \text{Trace}(\Sigma_1 - 2(\Sigma_1)^{\frac{1}{2}})$$

If $\Sigma = \text{diag}(\sigma_1^2, \ldots, \sigma_d^2)$ is diagonal, this reduces to:

$$W_2(q,p)^2 = ||\mu_1||^2 + \sum_{i=1}^{d}(\sigma_i - 1)^2$$

Forward KL-Divergence

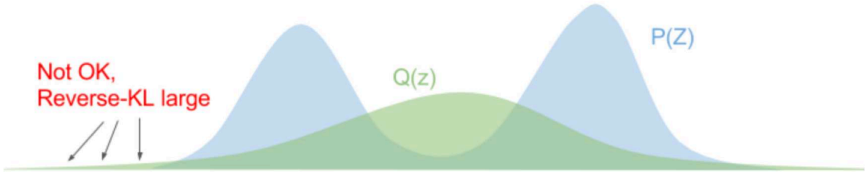
Given P, choose Gaussian Q to minimize $\mathrm{D}_{\mathrm{KL}}(P \parallel Q)$

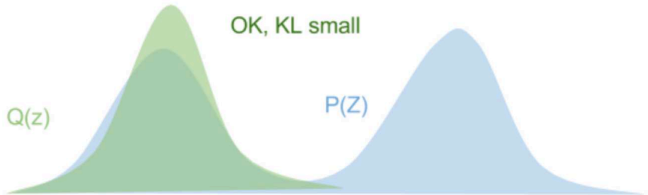


Q must **not** be small in **any** place where P is large.

Reverse KL-Divergence

Given P, choose Gaussian Q to minimize $\mathrm{D}_{\mathrm{KL}}(Q \parallel P)$





Q just needs to be concentrated in **some** place where P is large.