COMP9444: Neural Networks and Deep Learning

Week 2a. Probability

Raymond Louie

School of Computer Science and Engineering

Feb 25, 2025

Outline

1. Basic probability
   - ➢ 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)
2. Entropy and KL divergence
   - ➢ Entropy and KL-Divergence (3.13)
   - ➢ Wasserstein Distance

Probability is important!

1. Loss functions

2. Generative models (use probability distributions to generate new data)

3. Theoretical analysis of performance, e.g., universal approximators, bias/variance tradeoff

4. Hyperparameter tuning (probability-based optimization techniques)

5. And more..

Probability (3.1)

- Begin with a set Ω – the *sample space* (e.g. 6 possible rolls of a die) 1,2,3,4,5,6

  Each $\omega \in \Omega$ is a *sample point / possible world / atomic event* 1,2,3,4,5 and 6

- A *probability space* or *probability model* is a sample space with an assignment $P(\omega)$ for every $\omega \in \Omega$ such that

$$\boxed{\begin{gathered} 0 \leq P(\omega) \leq 1 \\ \sum_{\omega} P(\omega) = 1 \end{gathered}}$$

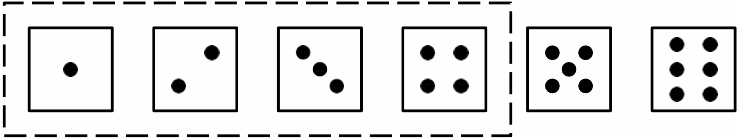
e.g. $P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = \frac{1}{6}$.

Random Events

A *random event* A is any subset of Ω

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

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



Question: What is P(A) where A = odd number?

Random Variables (3.2)

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

Example 1: Roll one dice. Outcome = 1,2,3,4,5,6. Function = odd. ,Range = True, False, True, etc..

Example 2: Roll two dice. Outcome = (1,1), (2,1), etc… Function = sum of numbers. Range = 2, 3, etc..

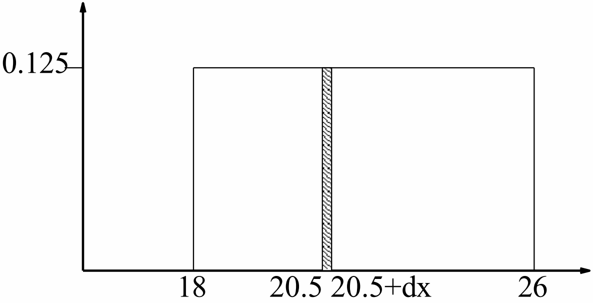
- P induces a *probability distribution* for any random variable X:

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

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

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



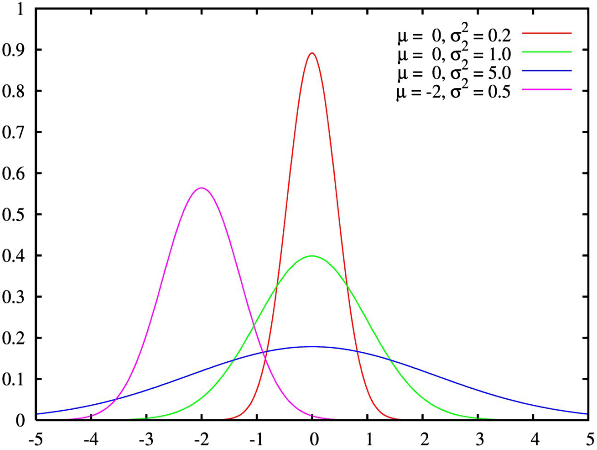
All intervals of same length are equally probable

Area of rectangle: 0.125*8=1

- 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)



Why isn't height the same?

μ = mean

σ = 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

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

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

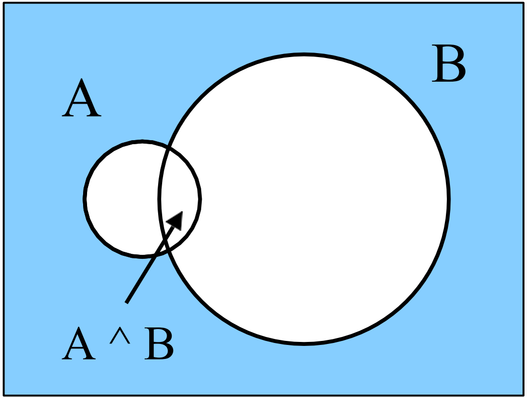
$$\boxed{P_{\mu,\Sigma}(x) = \prod_i P_{\mu_i,\sigma_i}(x_i)}$$

Why?

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

Probability and Logic- Example: not necessarily mutually exclusive

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

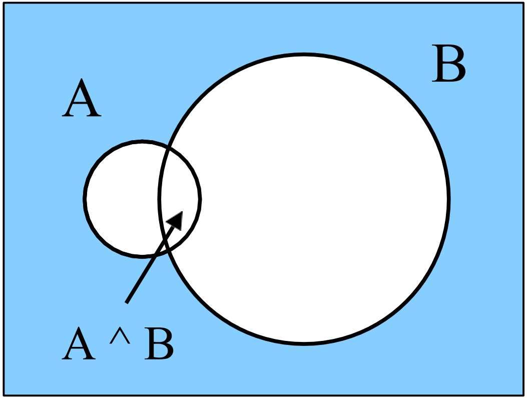


A: Odd, B: >3 (dice example)

P(Odd ∨ >3) = P(odd) + P(>3) – P(odd ∧ >3)
= ? + ? - ?

Probability and Logic- Example: not necessarily mutually exclusive

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



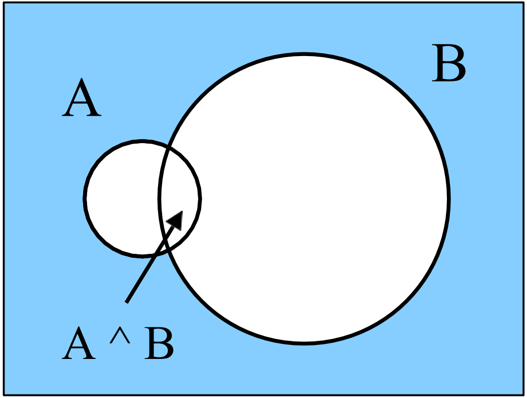
A: Odd, B: >3 (dice example)

P(Odd ∨ >3) = P(odd) + P(>3) – P(odd ∧ >3)
= P(1) + P(3) + P(5)
+ P(4) + P(5) + P(6)
- P(5)
= 3/6 + 3/6 – 1/6
= 5/6

Conditional Probability (3.5)

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

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

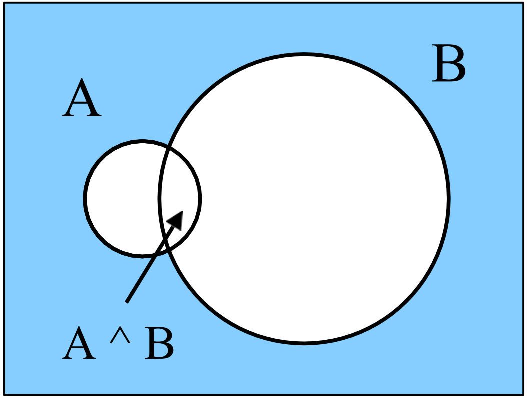


- A: Odd, B: >3 (dice example)
  - P(Odd | > 3) = ?
- A: Odd, B: <=3 (dice example)
  - P(Odd | <= 3) = ?

Conditional Probability (3.5)

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

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

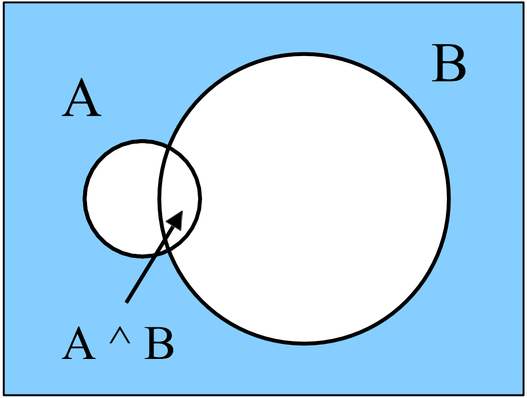


- A: Odd, B: >3 (dice example)
  - P(Odd and >3) = ?
  - P(>3) = ?
  - P(Odd | > 3) = 1/3
- A: Odd, B: <=3 (dice example)
  - P(Odd and <=3) = ?
  - P(<=3) = ?
  - P(Odd | <= 3) = 2/3

Conditional Probability (3.5)

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

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



- A: Odd, B: >3 (dice example)
  - P(Odd and >3) = P(5) = 1/6
  - P(>3) = P(4 or 5 or 6) = ½
  - P(Odd | > 3) = 1/6*2 = 1/3
- A: Odd, B: <=3 (dice example)
  - P(Odd and <=3) = P(1 or 3) = 2/6
  - P(<=3) = P(1 or 2 or 3) = ½
  - P(Odd | <= 3) = 2/6*2 = 2/3

Law of Total probability

$$\boxed{P(A) = \sum_n P(A \mid B_n) P(B_n)}$$

- A: Odd, B_1:>3, B_2 :<=3

- $P(A) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2)$
  $= (1/3)*(1/2) + (2/3)*(1/2)$
  $= 1/2$

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:

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

Typically calculated by law of total probability

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 (TP) of 98% and a specificity (TN) of 97%. If a patient tests positive, what is the probability that they have the cancer?

$$\boxed{P(\text{cancer} \mid \text{positive}) = \frac{P(\text{positive} \mid \text{cancer})P(\text{cancer})}{P(\text{positive})}}$$

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

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

$$P(\text{positive}) = P(\text{Positive} \mid \text{cancer})P(\text{cancer}) + P(\text{Positive} \mid \neg\text{cancer})P(\neg\text{cancer})$$

$$= \text{P}(\text{Positive} \mid \text{cancer})P(\text{cancer}) + (1- P(\text{negative} \mid \neg\text{cancer}))P(\neg\text{cancer})$$

$$= ? * ? + ? * ?$$

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 (TP) of 98% and a specificity (TN) of 97%. If a patient tests positive, what is the probability that they have the cancer?

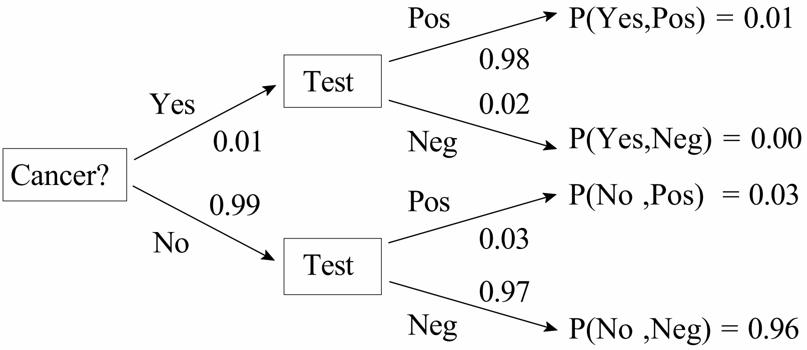
$$\boxed{P(\text{cancer} \mid \text{positive}) = \frac{P(\text{positive} \mid \text{cancer})P(\text{cancer})}{P(\text{positive})}}$$

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

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

$$\begin{aligned} P(\text{positive}) &= P(\text{Positive} \mid \text{cancer})P(\text{cancer}) + P(\text{Positive} \mid \neg\text{cancer})P(\neg\text{cancer}) \\ &= \text{P}(\text{Positive} \mid \text{cancer})P(\text{cancer}) + (1\text{-}\ P(\text{negative} \mid \neg\text{cancer}))P(\neg\text{cancer}) \\ &= 0.98 * 0.01 + 0.03 * 0.99 \end{aligned}$$

Bayes’ Rule for Medical Diagnosis



Cancer?

Yes 0.01

No 0.99

Test

Pos 0.98 → P(Yes,Pos) = 0.01

Neg 0.02 → P(Yes,Neg) = 0.00

Test

Pos 0.03 → P(No ,Pos) = 0.03

Neg 0.97 → P(No ,Neg) = 0.96

$$
\boxed{
\begin{aligned}
P(\text{cancer} \mid \text{positive}) &= \frac{P(\text{positive} \mid \text{cancer})P(\text{cancer})}{P(\text{positive})} \\
&= \frac{0.98 * 0.01}{0.98 * 0.01 + 0.03 * 0.99} = \frac{0.01}{0.01 + 0.03} = \frac{1}{4}
\end{aligned}
}
$$

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?

$$\boxed{P(\mathrm{A}\,|\,\mathrm{defect}) = \frac{P(\mathrm{defect}\,|\,\mathrm{A})P(\mathrm{A})}{P(\mathrm{defect})}}$$

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

In this case, the prior is: $P(A) = ?$, $P(B) = ?$

The conditional probabilities are:

$$P(\mathrm{defect}\,|\,\mathrm{A}) = ?, \quad \text{and} \quad P(\mathrm{defect}\,|\,\mathrm{B}) = ?$$

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?

$$\boxed{P(\mathrm{A}\,|\,\mathrm{defect}) \quad = \quad \frac{P(\mathrm{defect}\,|\,\mathrm{A})P(\mathrm{A})}{P(\mathrm{defect})}}$$

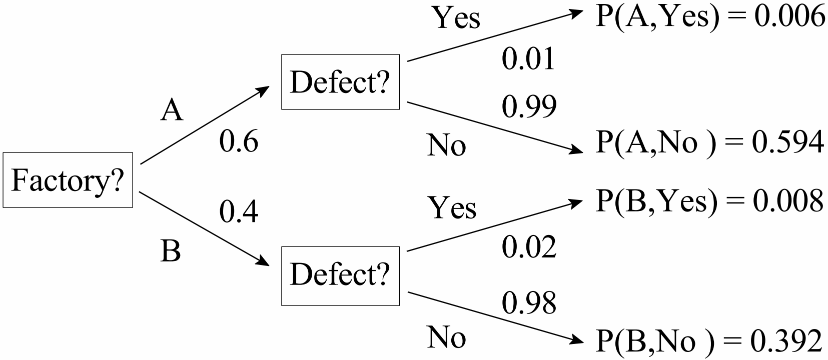
Answer: There are two random variables: Factory (A or B) and Defect (Yes or No).

In this case, the prior is: $P(A) = 0.6,$ $\qquad P(B) = 0.4$

The conditional probabilities are:

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

Bayes' Rule for Light Bulb Defects



$$
\begin{aligned}
P(\mathrm{A}\,|\,\mathrm{defect}) &= \frac{P(\mathrm{defect}\,|\,\mathrm{A})P(\mathrm{A})}{P(\mathrm{defect})} \\
&= \frac{0.01 * 0.6}{0.01 * 0.6 + 0.02 * 0.4} = \frac{0.006}{0.006 + 0.008} = \frac{3}{7}
\end{aligned}
$$

Outline

1. Basic probability
    - ➢ 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)
2. **Entropy and KL divergence**
    - ➢ Entropy and KL-Divergence (3.13)
    - ➢ Continuous Distributions
    - ➢ Wasserstein Distance

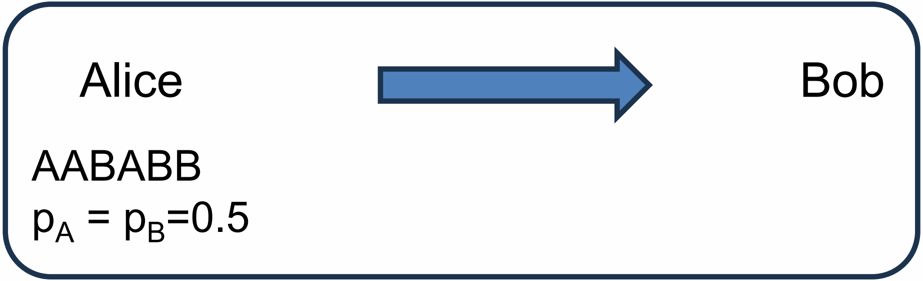
Motivation: loss functions between distributions

- In supervised machine learning, we want to minimize an error term, e.g., mean squared error

- It turns out that another metric, KL divergence maybe better in other situations (more on that later)

- **Idea:** Minimize the distribution of the output distribution and the trained distribution

Entropy Example 1



Alice

Bob

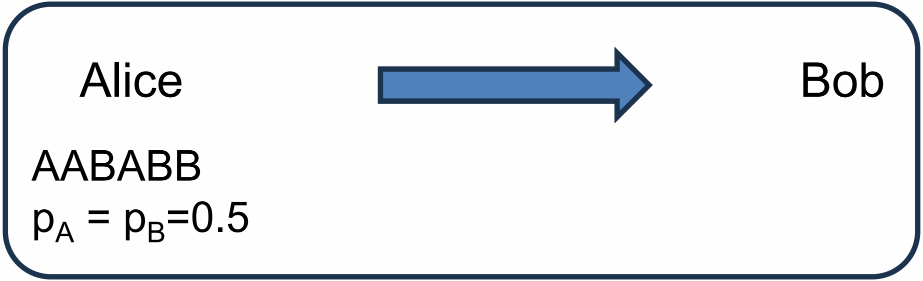
AABABB

$p_A = p_B = 0.5$

- Digital system, so need to binarize into 0 or 1
- What’s the more efficient coding scheme?

1) A=0, B=1: 0 0 1 0 1 1 (1 bit/symbol) or
2) A=00, B=11: 00 00 11 00 11 11 (2 bit/symbol)

Entropy Example 1



Alice

Bob

AABABB

$p_A = p_B = 0.5$

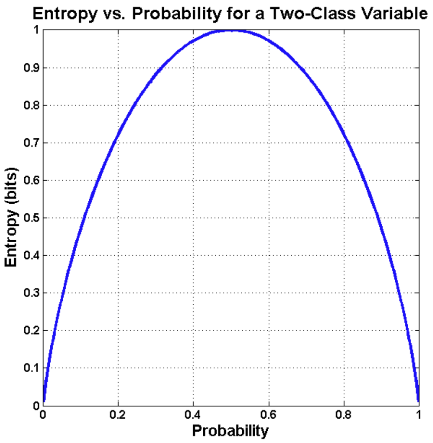
- The *entropy* of a discrete probability distribution $p = (p_1, \ldots, p_n)$ is

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

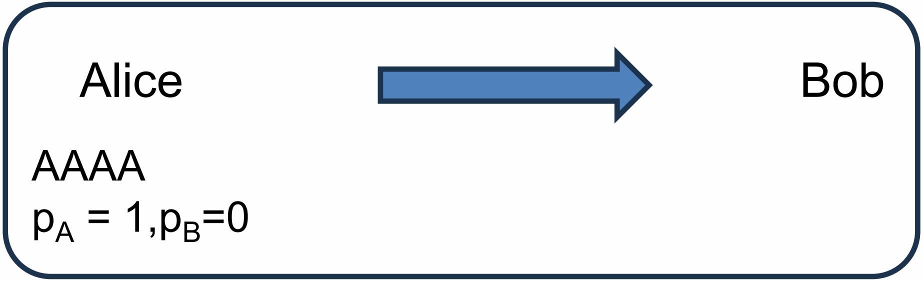
- 0.5(-log2(0.5) + 0.5(-log2(0.5)) = 1 bit

Entropy Alternative explanation

- Alternative interpretation: Measure of certainty (zero if 100% certain)
- Coin toss example



Entropy Example 1b



Alice

Bob

AAAA

$p_A = 1, p_B = 0$

- What is entropy?

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

Entropy and KL-Divergence (3.13)

- The *entropy* of a discrete probability distribution $p = (p_1, \ldots, p_n)$ is

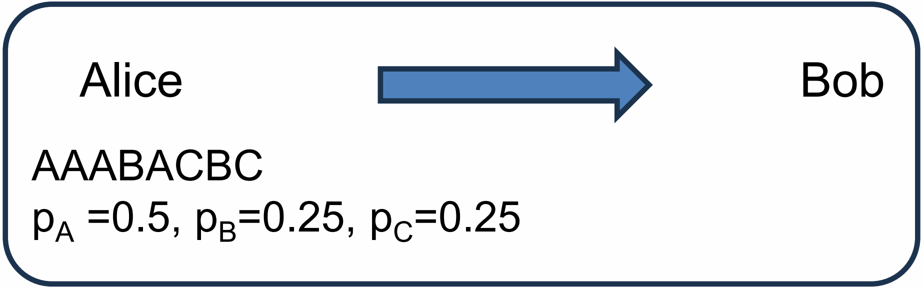
$$\boxed{H(p) = \sum_{i=1}^{n} p_i \left(-\log_2 p_i\right)}$$

- Entropy represents the minimum average number of bits needed to encode information from a probability distribution.

- Less bits> more certain, more bits-> less certain (outcomes are more spread out)

- Huffman coding uses this idea to assign variable-length codes to symbols such that frequently occurring symbols get shorter codes, minimizing the overall length of the encoded message.

Entropy example 2



Alice

Bob

AAABACBC

$p_A = 0.5$, $p_B = 0.25$, $p_C = 0.25$

$$H(p) = \sum_{i=1}^{n} p_i \left(-\log_2 p_i \right) = 0.5(-\log2(0.5)) + 0.25(-\log2(0.25)) + 0.25(-\log2(0.25)) = 1.5$$

- In this case, an optimally efficient code would be A=0, B=10, C=11.
- The average number of bits needed to encode information is:

$$0.5*1 + 0.25*2 + 0.25*2 = 1.5 \text{ bits/symbol}$$

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

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

- $\mathrm{D_{KL}}(q \,\|\, p)$ 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.

KL-Divergence (3.13)

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

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

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

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

Continuous Entropy and KL-Divergence

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

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

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

$$\boxed{\mathrm{D}_{\mathrm{KL}}(p \,\|\, q) = \int_{\theta} p(\theta)(\log\, p(\theta) - \log\, q(\theta))\, d\theta}$$

Entropy for Gaussian Distributions

- Entropy of Gaussian with mean μ and standard deviation σ:

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

Why do you expect entropy to increase with standard deviation?

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

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

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

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

KL-Divergence for Gaussians

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

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

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

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

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

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

How is KL divergence related to neural networks?

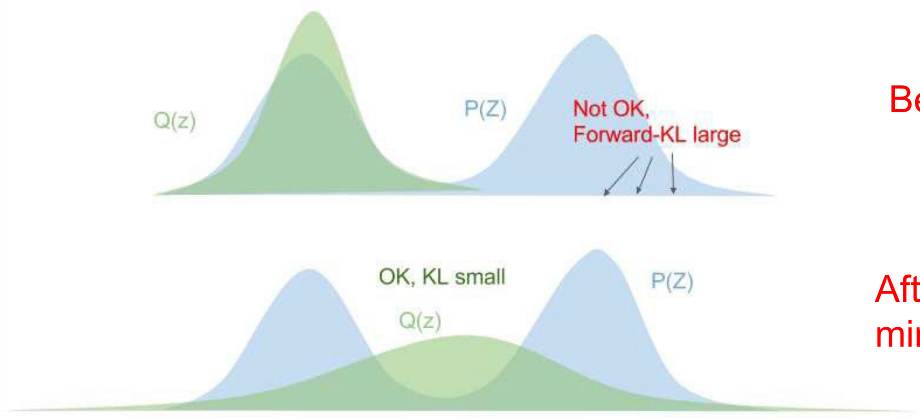
- P (true distribution) : distribution of the output from training data, e.g. P(cancer) vs P(healthy)

- Q (model distribution) : calculate from model output

- **Two ways:** Forward KL (D_{KL}(P || Q)) and Reverse KL-divergence (D_{KL} (Q || P))

Forward KL-Divergence (zero-avoiding)

Given P *(true distribution)* choose Gaussian Q *(model distribution)*to minimize $\mathrm{D}_{\mathrm{KL}}(P\,||\,Q)$



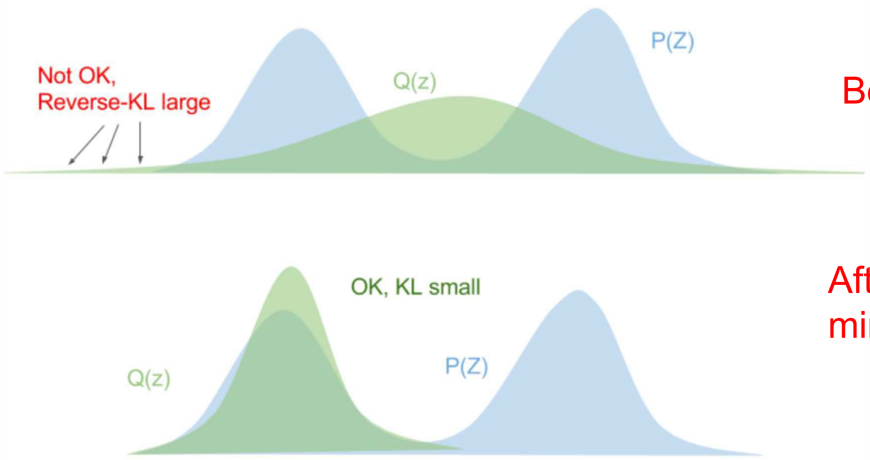
Before

After minimizing

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

Reverse KL-Divergence (zero-forcing)

Given P, choose Gaussian Q to minimize $\mathrm{D_{KL}}(Q \| P)$



Before

After minimizing

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

Wasserstein Distance

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

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

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

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

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

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

- **Why?** Vanishing gradients, etc.. https://medium.com/towards-data-science/why-wgans-beat-gans-a-journey-from-kl-divergence-to-wasserstein-loss-9ee5faf10b48