Fundamental Concepts of Generative Machine Learning

Erdem Akagündüz, PhD
Graduate School of Informatics, METU, Türkiye

ncc@ulakbim.gov.tr

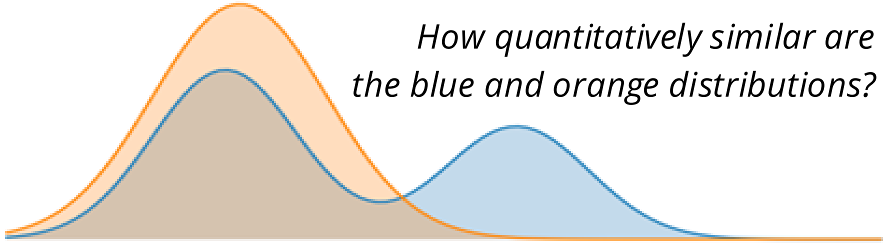
Lesson 1: Mathematical Background

Welcome to Part I: “Mathematical Background”

This part includes four subsections:

- Generation vs. Discrimination in Machine Learning
- Data Distribution, Sampling, Inference and Generation
- Expectation and Likelihood
- **Evaluation for Generative Models, Distribution Distances, Divergence and Entropy**

Distribution Distances



How quantitatively similar are the blue and orange distributions?

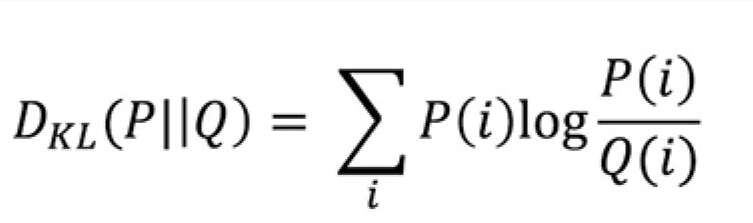
- Why do we need to compare distributions?
  - In generative modeling, it is important to compare different probability distributions to determine how well our model is performing.
  - For instance, we need to be able to evaluate the similarity between the true data distribution and the distribution learned by our generative model.
- “Divergence”
  - is a way to measure the distance between two probability distributions.
  - measures quantify how much one distribution differs from another in terms of their shapes, locations, or other characteristics.

Divergence (first)

- Not all distance measures between two distributions are “divergence” measures, but we will start with them first.
- Some of which, we may benefit from in this course, are:
  - Kullback-Leibler (KL) divergence
  - Jensen-Shannon (JS) divergence
  - Total Variation (TV) distance
  - Hellinger distance

Kullback-Leibler (KL) Divergence

- The KL divergence is a measure of the difference between two probability distributions, P and Q.
- It is defined as the expected value of the logarithmic difference between P and Q, where the expectation is taken with respect to P. The KL divergence is denoted as D(P||Q).



$$D_{KL}(P||Q) = \sum_{i} P(i) \log \frac{P(i)}{Q(i)}$$

Kullback-Leibler (KL) Divergence

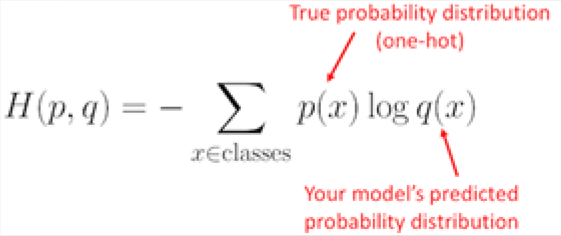
- The KL divergence is always non-negative, and it is zero if and only if the two distributions P and Q are identical.
- The KL divergence is not symmetric, meaning that D(P||Q) is not necessarily equal to D(Q||P).

$$D_{KL}(P||Q) \neq D_{KL}(P||Q)$$

- The KL divergence can be interpreted **as the amount of information lost when using Q to approximate P** (or vice versa). It measures the additional number of bits of information needed to specify P instead of Q.

Kullback-Leibler (KL) Divergence

- The KL divergence is commonly used in generative modeling to measure the similarity between the true data distribution and the distribution learned by a generative model.
- It is often used as a loss function to train generative models, such as Variational Autoencoders (VAEs), which aim to learn a lower-dimensional representation of the data that can be used to generate new samples.
- Did KL remind you of something?
    - Cross-entropy maybe?



True probability distribution (one-hot)

$$H(p,q) = -\sum_{x \in \text{classes}} p(x) \log q(x)$$

Your model's predicted probability distribution

KL vs CE

- Cross-entropy is a measure of the dissimilarity between two probability distributions, typically between a true distribution and an estimated distribution.

$$H(P,Q) = -\sum_{x} P(x)\log Q(x)$$

- KL divergence measures the divergence between two probability distributions, P and Q, by measuring the additional number of bits of information needed to specify P instead of Q.

$$KL(P|Q) = \sum_{x} P(x)\log \frac{P(x)}{Q(x)}$$

KL vs CE

- Both KL divergence and cross-entropy are used to measure the similarity or dissimilarity between two probability distributions.

$$H(P,Q) = -\sum_{x} P(x) \log Q(x)$$

- Both measures are commonly used in generative modeling to evaluate the performance of generative models and to optimize their parameters.

$$KL(P|Q) = \sum_{x} P(x) \log \frac{P(x)}{Q(x)}$$

- Both measures are non-negative and minimize to zero if and only if the two distributions are identical.
- Both measures are asymmetrical.

KL vs CE

- KL Divergence focuses on the additional information needed to be accurate about the true distribution when starting with an approximation. It emphasizes the "gap" between the true distribution and the approximation.
- Cross Entropy is more about the efficiency of encoding events from the true distribution when using the code optimized for an approximation. It looks at the average number of bits needed and emphasizes the cost of using the code optimized for

$$H(P,Q) = -\sum_{x} P(x) \log Q(x)$$

$$KL(P|Q) = \sum_{x} P(x) \log \frac{P(x)}{Q(x)}$$

Jensen-Shannon (JS) Divergence

- The JS divergence is a symmetric measure of the difference between two probability distributions.
- It is a smoothed version of the KL divergence, which can be used to compare two probability distributions that may have disjoint support.

$$D_{JS}(p||q) = \frac{1}{2} D_{KL}(p||\frac{p+q}{2}) + \frac{1}{2} D_{KL}(q||\frac{p+q}{2})$$

Jensen-Shannon (JS) Divergence

- The JS divergence is a symmetric measure of the difference between two probability distributions.
- It is a smoothed version of the KL divergence, which can be used to compare two probability distributions that may have disjoint support.
- The JS divergence is bounded by 0 and 1, and is equal to 0 if and only if the two distributions are identical.

$$0 \leq \mathrm{JSD}(P \parallel Q) \leq 1$$

- The JS divergence are used as a loss function to train generative models, such as Generative Adversarial Networks (GANs).

Information Theory

- Information theory was initially developed to understand and improve communication systems, especially in the context of telegraphy and radio.
- However, its scope has expanded to various other areas such as data compression, cryptography, and more recently, deep learning and generative models.
- In the context of deep generative models, information theory provides a theoretical framework for understanding and designing models that can generate high-quality and diverse samples from complex distributions.

Entropy & Information

- Entropy is a measure of uncertainty or disorder in a random variable, while information is the reduction of uncertainty or surprise gained from an event.
- In deep generative models,
  - the entropy of the output distribution can be used to measure the complexity of the generated samples,
  - while information can be used to measure the amount of information captured in the learned representation.
- **<u>They are formulated measures!</u>**

Shannon’s Entropy

- Introduced by Claude Shannon in 1948 as a measure of uncertainty or information content in a random variable or probability distribution, defined as the expected value of the information contained in each possible outcome, given the probability distribution:

$$H(X) = -\sum_{i} P(x_i) \log P(x_i)$$

- maximized when all outcomes are equally likely (i.e., maximum uncertainty)
- minimized when there is only one possible outcome (i.e., no uncertainty)

Conditional Entropy

- is the amount of uncertainty remaining in a random variable given that another random variable has been observed or known.

$$H(X) = -\sum_{i} P(x_i) \log P(x_i) \longrightarrow \mathrm{H}(Y|X) = -\sum_{x\in\mathcal{X},y\in\mathcal{Y}} p(x,y) \log \frac{p(x,y)}{p(x)}$$

- Conditional entropy tells us how much information we gain about Y by observing X. If conditional entropy is high, it means that observing X gives us little information about Y, and vice versa.

Entropy and Generative Models

- Entropy measures such as Shannon entropy and differential entropy are used to quantify the uncertainty or randomness in the generated samples.
- In a well-trained generative model, the generated samples should have high entropy, indicating that the model is able to produce a diverse set of samples that capture the variability in the training data.
    - Example: In a generative adversarial network (GAN), the generator tries to generate samples that fool the discriminator. The entropy of the generated samples can be used to measure the diversity and quality of the samples generated by the GAN.

Shannon’s “Self-Information”

- The amount of information gained by an event with probability p is defined as:

$$\mathrm{I}(x) := -\log_b\left[\Pr\left(x\right)\right] = -\log_b\left(P\right).$$

- Shannon's definition of self-information meets several axioms:
  - An event with probability 100% is perfectly unsurprising and yields no information.
  - The less probable an event is, the more surprising it is and the more information it yields.
  - If two independent events are measured separately, the total amount of information is the sum of the self-informations of the individual events

Mutual Information

- is a measure of the amount of information that two variables share.
- MI quantifies the reduction in uncertainty about one variable given knowledge of the other variable, and can be represented using the Entropy:

$$\begin{aligned} \mathrm{I}(X;Y) &\equiv \mathrm{H}(X) - \mathrm{H}(X \mid Y) \\ &\equiv \mathrm{H}(Y) - \mathrm{H}(Y \mid X) \\ &\equiv \mathrm{H}(X) + \mathrm{H}(Y) - \mathrm{H}(X,Y) \\ &\equiv \mathrm{H}(X,Y) - \mathrm{H}(X \mid Y) - \mathrm{H}(Y \mid X) \end{aligned}$$

  - Example: MI can be used as a regularizer to encourage disentanglement of the latent variables in the learned representation.

Information and Generative Models

- Information measures such as mutual information and conditional entropy are used to evaluate the ability of the generative model to capture the underlying structure of the data.
- In a well-trained generative model, the mutual information between the generated samples and the training data should be low, indicating that the generated samples are not duplicating the training data.
  - Example: In a variational autoencoder (VAE), the encoder tries to compress the input data into a low-dimensional latent space. The mutual information between the latent space and the input data can be used to measure the amount of information that is preserved in the latent space (like a regularizer to encourage disentanglement of the latent variables in the learned representation)

Next lecture:

- **PART II: “Latent Spaces”**

- **(The Curse of) Dimensionality, Deep Features vs. Latent Spaces**
- Latent Space properties, Continuity, Entanglement, etc

Drive thru: Will that be all?
Me:
The First Order was only the beginning!

Thanks

This project has received funding from the European High-Performance Computing Joint Undertaking (JU) under grant agreement No 101101903. The JU receives support from the Digital Europe Programme and Germany, Bulgaria, Austria, Croatia, Cyprus, Czech Republic, Denmark, Estonia, Finland, Greece, Hungary, Ireland, Italy, Lithuania, Latvia, Poland, Portugal, Romania, Slovenia, Spain, Sweden, France, Netherlands, Belgium, Luxembourg, Slovakia, Norway, Türkiye, Republic of North Macedonia, Iceland, Montenegro, Serbia