Basics of Information Theory

Mohammad Reza Karimi

Spring 2020

Entropy

Let $p = (p_1, \ldots, p_k)$ be a distribution over k objects. The **entropy** of p is defined as

$$H(p) = -\sum_{i=1}^{k} p_i \log p_i.$$

There are several intuitions about (Shannon's) entropy, most importantly the **compression** idea, but we describe another interesting one.

Let $X_1, \ldots, X_n$ be independent draws from the distribution p. Define $Y_i = -\log p(X_i)$. It is easy to check that $\mathbb{E}[Y_i] = H(p)$. We now use the law of large numbers:

$$\forall \varepsilon > 0, \quad \lim_{n \to \infty} \mathbb{P}\left(\left| \frac{1}{n} \sum_{i=1}^{n} Y_i - H(p) \right| \leq \varepsilon \right) = 1.$$

But we know that

$$\frac{1}{n}\sum_{i=1}^{n} Y_i = -\frac{1}{n}\log p(X_1, \ldots, X_n),$$

resulting in

$$2^{-n(H+\varepsilon)} \leq p(X_1, \ldots, X_n) \leq 2^{-n(H-\varepsilon)},$$

with high probability!

Another way to state this is that

> For large n, the distribution over sequences is like a uniform distribution over Group A, and zero on Group B.

We call sequences in Group A, the **"typical sequences"**.

As an example, take a biased coin with distribution $(\frac{3}{4}, \frac{1}{4})$. We toss this coin 1000 times. Now the "most probable outcome" of this experiment, is the sequence of all heads. For this sequence we have $\frac{1}{n}\sum Y_i = -\frac{1}{1000}\log(\frac{3}{4})^{1000} \approx 0.125$, but the entropy of the distribution is $-\frac{3}{4}\log\frac{3}{4} - \frac{1}{4}\log\frac{1}{4} \approx 0.811$. These two numbers are far away, and this makes the "all heads" sequence not a typical sequence.

One thing to keep in mind, is that the size of the typical set is determined by entropy. The higher the entropy, the larger the set of typical sequences. In the extreme case, if we have an unbiased coin, then every sequence would be typical. By simple calculation, one can show that the size of typical set is approximatly equal to $2^{nH(p)}$, so

Entropy is a measure of the *volume of the typical set*,

another intuition!

Kullback-Leibler Divergence

Let x be a string of length n over the alphabet $\{1, \ldots, k\}$. For x we can compute the frequencies of each alphabet letter and put all these frequencies in a vector $(p_1, \ldots, p_k)$. We call this vector the "type of x" and write it as P_x. For example, if $x = 1314231$, and the alphabet is $\{1, \ldots, 4\}$, then $P_x = (\frac{3}{7}, \frac{1}{7}, \frac{2}{7}, \frac{1}{7})$.

Exercise 1. *Show that the number of sequences having a certain type P, is approximately equal to $2^{-nH(P)}$.*

Now take an arbitrary distribution Q over the alphabet. The question that we can ask now is what is the probability of observing a sequence of type P under the assumption that the sequence is generated by Q. Interestingly we can compute this probability and in the limit ($n \to \infty$) the solution would be approximately equal to...

$$2^{-n\mathrm{KL}(P\|Q)},$$

where

$$\mathrm{KL}(P\|Q) = \sum_{i=1}^{k} p_i \log \frac{p_i}{q_i},$$

is called the Kullback-Leibler Divergence of P from Q.

As an example, let us say that we toss a fair coin 1000 times, and ask what is the probability to get a sequence in which $\frac{3}{4}$ of the outcomes are heads and $\frac{1}{4}$ of the outcomes are tails. We can see that

$$\mathrm{KL}((\tfrac{3}{4},\tfrac{1}{4})\|(\tfrac{1}{2},\tfrac{1}{2})) = \frac{3}{4}\log\frac{3}{2} + \frac{1}{4}\log\frac{1}{2} \approx 0.189,$$

and by the result above we see that the probability is about $2^{-189} \approx 10^{-57}$.

This means that the higher the divergence of the "candidate distribution" P and the "true distribution" Q gets, the lower would be the probability of observing an outcome of P. That is why we can see the KL divergence as a measure of "distance" between distributions.

Properties of KL Divergence

- Always $\mathrm{KL}(P\|Q) \geq 0$.
- For product distributions, KL is additive. That is,

$$\mathrm{KL}(P_1 \otimes P_2 \| Q_1 \otimes Q_2) = \mathrm{KL}(P_1\|Q_1) + \mathrm{KL}(P_2\|Q_2).$$

- The Pinsker Inequality:

$$d_{\mathrm{TV}}(P,Q)^2 \leq 2\,\mathrm{KL}(P\|Q)$$

Application: Testing a Coin

We are given a coin, but we don't know if it is biased or not. We only know that the bias is either $1/2$ or $1/2 + \epsilon$.

Question 1. *How many times we should toss this coin, so that we can tell if it is the biased coin or not?*

Let us denote by $X = (X_1, \ldots, X_n)$ the results of the tosses. Suppose that we have a decision rule ψ that

$$\psi : X \mapsto \{B, U\}.$$

Then, the probability of ψ making a mistake is

$$\mathbb{P}[\text{error}] = \frac{1}{2}\,\mathbb{P}_B[\psi(X) \neq B] + \frac{1}{2}\,\mathbb{P}_U[\psi(X) \neq U].$$

The following theorem tells us what is the best we can achieve:

Theorem 1. *We have*

$$\inf_{\psi}\{\mathbb{P}_B[\psi(X) \neq B] + \mathbb{P}_U[\psi(X) \neq U]\} = 1 - d_{\mathrm{TV}}(\mathbb{P}_B, \mathbb{P}_U).$$

Proof. Let $A \subset \Omega$ be the set $\{X : \psi(X) = B\}$. Note that a classifier is identified by its acceptance set A. We have

$$\begin{aligned}\mathbb{P}_B[\psi(X) \neq B] + \mathbb{P}_U[\psi(X) \neq U] &= \mathbb{P}_B(A^c) + \mathbb{P}_U(A) \\ &= 1 - \mathbb{P}_B(A) + \mathbb{P}_U(A).\end{aligned}$$

Taking the infimum gives

$$\begin{aligned}\inf_{\psi}\{\cdots\} &= \inf_{A}\{1 - \mathbb{P}_B(A) + \mathbb{P}_U(A)\} \\ &= 1 - \sup_{A}(\mathbb{P}_B(A) + \mathbb{P}_U(A)) \\ &= 1 - d_{\mathrm{TV}}(\mathbb{P}_B, \mathbb{P}_U).\end{aligned}$$

□

Now using Pinsker inequality, we understand that the least probability of error is bounded below by

$$1 - \sqrt{2\,\mathrm{KL}(\mathbb{P}_B \,\|\, \mathbb{P}_U)}.$$

Remains to compute the KL divergence. Note that both $\mathbb{P}_B$ and $\mathbb{P}_U$ are product probability distributions.

Denote by p_B the distribution of a single biased coin and p_U likewise. Then

$$\begin{aligned}
\mathrm{KL}(\mathbb{P}_B \,\|\, \mathbb{P}_U) &= n \cdot \mathrm{KL}(p_B \| p_U) \\
&= n \cdot \left((\tfrac{1}{2} + \epsilon) \log \frac{\frac{1}{2} + \epsilon}{\frac{1}{2}} + (\tfrac{1}{2} - \epsilon) \log \frac{\frac{1}{2} - \epsilon}{\frac{1}{2}} \right) \\
&\approx n \cdot \big((\tfrac{1}{2} + \epsilon)(1 + 2\epsilon) + (\tfrac{1}{2} - \epsilon)(1 - 2\epsilon) \big) \\
&= n \cdot (1 + 4\epsilon^2) = O(n\epsilon^2)
\end{aligned}$$

So for example, if we want to have $\mathbb{P}[\text{error}] \leq 1/4$, we should have

$$n = \Omega(\epsilon^{-2}).$$

Information theory in loss functions

Mohammad Reza Karimi

Anastasia Makarova

Spring 2020

Part 1:
Basics of Information Theory

Entropy

Let $p = (p_1, \ldots, p_k)$ be a distribution over k objects. The **entropy** of p is defined as

$$H(p) = -\sum_{i=1}^{k} p_i \log p_i.$$

There are several intuitions about (Shannon's) entropy, most importantly the **compression** idea, but we describe another interesting one.

Let $X_1, \ldots, X_n$ be independent draws from the distribution p. Define $Y_i = -\log p(X_i)$. It is easy to check that $Y_i = H(p)$. We now use the law of large numbers:

$$\forall \varepsilon > 0, \quad \lim_{n \to \infty} ¶\left(\left| \frac{1}{n} \sum_{i=1}^{n} Y_i - H(p) \right| \leq \varepsilon \right) = 1.$$

But we know that

$$\frac{1}{n}\sum_{i=1}^{n} Y_i = -\frac{1}{n}\log p(X_1, \ldots, X_n),$$

resulting in

$$2^{-n(H+\varepsilon)} \leq p(X_1, \ldots, X_n) \leq 2^{-n(H-\varepsilon)},$$

with high probability!

Another way to state this is that

> For large n, the distribution over sequences is like a uniform distribution over Group A, and zero on Group B.

We call sequences in Group A, the **"typical sequences"**.

As an example, take a biased coin with distribution $(\frac{3}{4}, \frac{1}{4})$. We toss this coin 1000 times. Now the "most probable outcome" of this experiment is

For this sequence we have $\frac{1}{n}\sum Y_i =$

but the entropy of the distribution is $-\frac{3}{4}\log\frac{3}{4} - \frac{1}{4}\log\frac{1}{4} \approx 0.811$.

One thing to keep in mind, is that the size of the typical set is determined by entropy.

By simple calculation, one can show that (**Exercise**) the size of typical set is approximatly equal to $2^{nH(p)}$, so

Entropy is a measure of the *volume of the typical set*,

another intuition to keep in mind!

Kullback-Leibler Divergence

Let x be a string of length n over the alphabet $\{1, \ldots, k\}$. For x we can compute the frequencies of each alphabet letter and put all these frequencies in a vector $(p_1, \ldots, p_k)$. We call this vector the **type of** x and write it as P_x.

Exercise 1. *Show that the number of sequences having a certain type P, is approximately equal to $2^{-nH(P)}$.*

Now take an arbitrary distribution Q over the alphabet.

Question 1. *What is the probability of observing a sequence of type P under the assumption that the sequence is generated by Q.*

Interestingly we can compute this probability and in the limit ($n \to \infty$) the solution would be approximately equal to...

$$2^{-nPQ}$$

where

$$PQ = \sum_{i=1}^{k} p_i \log \frac{p_i}{q_i},$$

is called the Kullback-Leibler Divergence of P from Q.

As an example, let us say that we toss a fair coin 1000 times, and ask what is the probability to get a sequence in which $\frac{3}{4}$ of the outcomes are heads and $\frac{1}{4}$ of the outcomes are tails. We can see that

$$(\tfrac{3}{4}, \tfrac{1}{4})(\tfrac{1}{2}, \tfrac{1}{2}) = \qquad ,$$

and by the result above we see that the probability is about

This means that the higher the divergence of the "candidate distribution" P from the "true distribution" Q gets, the lower would be the probability of observing an outcome of P. That is why we can see the KL divergence as a measure of "distance" between distributions.

Properties of KL Divergence

- Always $PQ \geq 0$.
- For product distributions, KL is additive. That is,

$$P_1 \otimes P_2 Q_1 \otimes Q_2 = P_1 Q_1 + P_2 Q_2.$$

- The Pinsker Inequality:

$$P, Q^2 \leq 2\, PQ$$

Application: Testing a Coin

We are given a coin, but we don't know if it is fair or not. We only know that the bias is either $1/2$ or $1/2 + \epsilon$.

Question 2. *How many times we should toss this coin, so that we can tell if it is the biased coin or not?*

Let us denote by $X = (X_1, \ldots, X_n)$ the results of the tosses. Suppose that we have a decision rule ψ that

$$\psi : X \mapsto \{B, U\}.$$

Then, the probability of ψ making a mistake is

$$\P[\text{error}] = \frac{1}{2}\P_B\psi(X) \neq B + \frac{1}{2}\P_U\psi(X) \neq U.$$

The following theorem tells us what is the best we can achieve:

Theorem 1. *We have*

$$\inf_{\psi} *\P_B\psi(X) \neq B + \P_U\psi(X) \neq U = 1 - \P_B, \P_U.$$

Proof. Let $A \subset \Omega$ be the set $X : \psi(X) = B$.

Taking the infimum gives

□

Now we use Pinsker inequality:

Denote by p_B the distribution of a single biased coin and p_U likewise.

Then

$$\P_B \P_U = n \cdot p_B p_U$$

So for example, if we want to have ¶[error] $\leq 1/4$, we should have

$$n = \Omega(\epsilon^{-2}).$$

Learned Concepts

1. Shannon Entropy
2. Typical Sequence
3. Kullback-Leibler Divergence
4. Total Variation Distance

After break

1. Cross-entropy Loss (CE)
2. Relation to MLE and KL
3. Demo for intuition
4. Handling imbalance with CE

Maximum Likelihood Estimation (Recap)

Predicted likelihood $\hat{P}(Y, X|\mathbf{w})$

Data $D = (Y, X) = \{(x_i, y_i)\}_{i=1}^{n}$

w model parameters

MLE $\underset{w}{\operatorname{argmax}}\, \hat{P}(Y, X|w) = \underset{w}{\operatorname{argmax}}\, \hat{P}(Y|X, w) \overset{iid}{=}$

$= \underset{w}{\operatorname{argmax}} \prod_{i=1}^{n} \hat{P}(y_i|x_i, w) =$

$= \underset{w}{\operatorname{argmin}} - \sum_{i=1}^{n} \log P(y_i|x_i, w)$ (*)

MLE and Cross-entropy

discrete contin

- Def Cross-entropy $H(P,Q) = \mathbb{E}_P[-\log Q] = -\sum_{i=1}^{n} p_i \log q_i = -\int_x p(x) \log q(x)$

- $p = P(Y|x)$ true distribution
  $\hat{p}_w = \hat{P}(Y|x,w)$ predictive dist

$$P(Y|x_i) = \begin{cases} 1 & \text{if } y = y_i \\ 0 & \text{else} \end{cases}$$

$\equiv$ one-hot encoding

$$\sum_{c \in C} \hat{P}(y=c|x,w) = 1$$

- Cross-entropy: for point x_i

$$H_i(p, \hat{p}_w) = -\sum_{y \in Y} p(y|x_i) \log \hat{p}_w(y|x_i) = -\log \hat{p}_w(y_i|x_i)$$

For dataset D

$$L = -\sum_{i=1}^{n} \log p_w(y_i|x_i) \rightarrow \min_w$$

same as (*)

KL divergence and Cross-entropy

- $KL(P \| \hat{P}_w) = \mathbb{E}_P[-\log \frac{\hat{P}_w}{P}] = \mathbb{E}_P[-\log \hat{P}_w + \log P] = \underbrace{\mathbb{E}_P[-\log \hat{P}_w]}_{H(P, \hat{P}_w)} - \underbrace{\mathbb{E}_P[-\log P]}_{H(P)}$

$$H(P, P_w) = H(P) + KL(P \| P_w) \xrightarrow[w]{\min} \equiv$$

min CE is exactly what KL min

- Is $H(P, P_w)$ the same as reversed $H(P_w, P) = \underbrace{KL(P_w \| P)}_{:= \text{reversed KL}} + H(P_w)$ for classification?

$P(y|x_i)$ is one-hot encoding where all mass is on $y = y_i$

$H(P, \hat{P}_w)$ computed only in points where $P(y=y_i|x_i) = 1$ and $\hat{P}_w$ is ignored in other points $y \in \mathcal{Y}$.

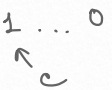
Alternatively, $H(\hat{P}_w, P)$ doesn't penalize for points where $P(y|x) > 0$, but $\hat{P}_w(y|x) = 0$.

(1)

No

What will happen with CE loss for the data point x, if your model predicts class c with 0 probability, while it actually pops up in reality? Specifically,

$$\hat{p}(y = c|x, w) = 0, y_{true} = c \longrightarrow$$



- CE(x) will be infinitely small
- CE(x) will be infinitely big

$$CE(x) = -\sum_{y \in Y} p(y|x) \cdot \log \hat{p}(y_i = y|x, w) =$$

0

Eduapp

What will happen with CE loss for the data point x, if your model predicts class c with 0 probability, while it actually pops up in reality? Specifically,

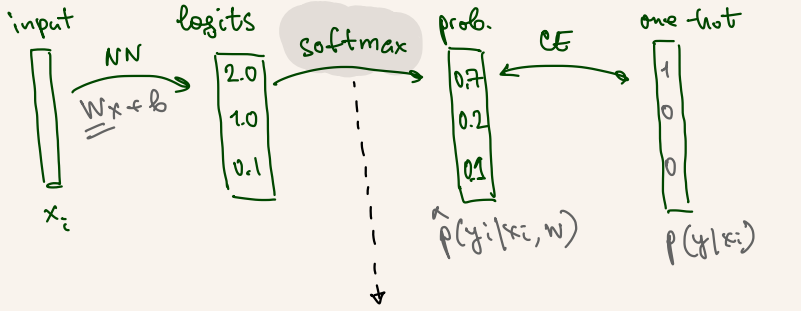
$$\hat{p}(y = c|x, w) = 0, y_{true} = c$$

- CE(x) will be infinitely small
- CE(x) will be infinitely big

Eduapp

How to avoid that surprise?

CE into the wild (NN case)



Softmax functions as hypothesis models are conservative enough

Why CE?

3 class classification cat / dog / rabbit

Model 1 output

output $\hat{p}_w$			target			correct?
0.3	0.3	0.4	0	0	1	1
0.3	0.4	0.3	0	1	0	1
0.2	0.1	0.7	1	0	0	0

classification error = 1/3

misclassification error is the same

$CE_1 = -(0 \cdot \log(0.3) + \dots + \log(0.4))$

$L = \frac{1}{n}\sum_{i=1}^{n} CE_i = \frac{1}{3}(\log 0.4 + \log 1.4 + \log(0.7)) = 1.38$

Model 2 output

output $\hat{p}_w$			target			correct?
0.1	0.2	0.7	0	0	1	1
0.1	0.7	0.2	0	1	0	1
0.3	0.4	0.3	1	0	0	0

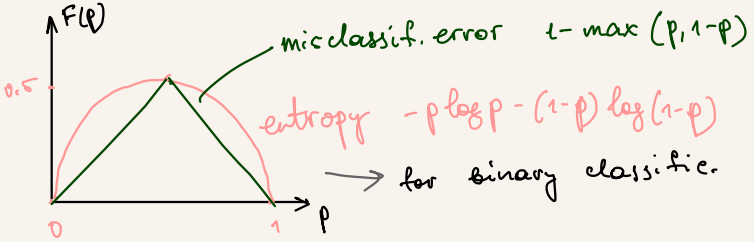
classif. error = 1/3

$CE_1 = -(0 \cdot \log(0.1) + 0 \cdot \log(0.2) + 1 \cdot \log(0.7$

$L = \frac{1}{n}\sum_{i=1}^{n} CE_i = \frac{1}{3}$

$= 0.64$

Why CE?



1. CE is more sensitive to p
2. differentiable

? Why not MSE ? – hw: compare gradient step in hw5 for CE with gradient step for MSE

Visualization

https://www.desmos.com/calculator/zytm2sf56e

Here p is continuous, though in classification it is discrete

Try different configurations:

(i) p is unimode (ii) p has 2 modes (iii) $\mathcal{H}(p,q) \neq \mathcal{H}(p,q)$

to see how CE & KL change

Weighted CE loss

What if $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ is imbalanced ?

\# binary classification : logistic

$|\mathcal{D}_1| = \frac{n}{10}$ and $|\mathcal{D}_2| = \frac{9n}{10}$

$$\nabla_w \ell_i(w) = P(y_i \neq y | x) \cdot (-yx)$$

$$\nabla_w L(w) = \sum_{i=1}^{n} \nabla_w \ell_i(w)$$

use weighted CE loss to deal with imbalance

Learned Concepts

1. Basics of Information Theory
2. KL divergence and Cross-Entropy
3. Imbalance in data

Eduapp: the question