Data Science

Sparse_categorical_crossentropy vs categorical_crossentropy (keras, accuracy)

Asked 3 years, 4 months ago Modified 1 year, 2 months ago Viewed 47k times

59

Which is better for accuracy or are they the same? Of course, if you use `categorical_crossentropy` you use one hot encoding, and if you use `sparse_categorical_crossentropy` you encode as normal integers. Additionally, when is one better than the other?

neural-network keras loss-function encoding

30

Share Edit Follow

edited Feb 7, 2021 at 21:05
Ethan
1,391 ● 8 ● 17 ● 37

asked Dec 1, 2018 at 6:28
Master M
693 ● 1 ● 6 ● 5

2 Answers

Sorted by: Highest score (default)

59

✓

Use sparse categorical crossentropy when your classes are mutually exclusive (e.g. when each sample belongs exactly to one class) and categorical crossentropy when one sample can have multiple classes or labels are soft probabilities (like [0.5, 0.3, 0.2]).

Formula for categorical crossentropy (S - samples, C - classess, $s \in c$ - sample belongs to class c) is:

$$-\frac{1}{N}\sum_{s\in S}\sum_{c\in C}1_{s\in c}logp(s\in c)$$

For case when classes are exclusive, you don't need to sum over them - for each sample only non-zero value is just $-logp(s \in c)$ for true class c.

This allows to conserve time and memory. Consider case of 10000 classes when they are mutually exclusive - just 1 log instead of summing up 10000 for each sample, just one integer instead of 10000 floats.

Formula is the same in both cases, so no impact on accuracy should be there.

Share Edit Follow

edited Sep 14, 2019 at 13:54

answered Dec 1, 2018 at 8:20
featuredpeow
706 ● 6 ● 4

1 Do they impact the accuracy differently, for example on mnist digits dataset? – Master M Dec 1, 2018 at 8:47

1 Mathematically there is no difference. If there is significant difference in values computed by implementations (say tensorflow or pytorch), then this sounds like a bug. Simple comparison on random data (1000 classes, 10 000 samples) show no difference. – featuredpeow Dec 1, 2018 at 14:20

Dear frenzykryger, I guess you forgot a minus for the one sample case only: "for each sample only non-zero value is just -log(p(s ∈ c))". For the rest, nice answer. – Nicg Sep 13, 2019 at 12:48

You're right. Thanks! – featuredpeow Sep 14, 2019 at 13:54

@frenzykryger I am working on multi-output problem. I have 3 seperate output `o1,o2,o3` and each one have `167,11,7` classes respectively. I've read your answer that it'll make no difference but is there any difference if I use `sparse__` or not. Can I go for `categorical` for the last 2 and `sparse` for the first one as there are 167 classes in the first class? – Deshwal Jan 8, 2020 at 4:58