Long-tail distribution

Peng YUN

20200818

1 A general description of the long-tail distribution

The natural occurrence of classes in the world is a long tailed distribution [1, 2, 3], whereby instances for most classes are rare and instances for few classes are abundant.

2 Another general description of the long-tail distribution

"It should be noted, however, that even when one has an apparently massive data set, the effective number of data points for certain cases of interest might be quite small. In fact, data across a variety of domains exhibits a property known as the **long tail**, which means that a few things (e.g. words) are very common, but most things are quite rare. For example, 20% of Google searches each day have never been seen before[4]."

References

[1] Xiangxin Zhu, Dragomir Anguelov, and Deva Ramanan. Capturing long-tail distributions of object subcategories. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[2] Grant Van Horn and Pietro Perona. The devil is in the tails: Fine-grained classification in the wild. *arXiv preprint arXiv:1709.01450*, 2017.

[3] Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. Learning to Model the Tail. In I Guyon, U V Luxburg, S Bengio, H Wallach, R Fergus, S Vishwanathan, and R Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 7029–7039. Curran Associates, Inc., 2017.

[4] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.