Title: Perplexity

URL: https://en.wikipedia.org/wiki/Perplexity

PageID: 4631023

Categories: Category:Entropy and information, Category:Language modeling

---

In information theory , perplexity is a measure of uncertainty in the value of a sample from a discrete probability distribution . The larger the perplexity, the less likely it is that an observer can guess the value which will be drawn from the distribution. Perplexity was originally introduced in 1977 in the context of speech recognition by Frederick Jelinek , Robert Leroy Mercer , Lalit R. Bahl, and James K. Baker . [1]

Perplexity of a probability distribution

The perplexity PP of a discrete probability distribution p is a concept widely used in information theory, machine learning , and statistical modeling. It is defined as

P P (p) = ∏ x p (x) − p (x) = b − ∑ x p (x) log b ■ p (x) {\displaystyle {\mathit {PP}}(p)=\prod _{x}p(x)^{-p(x)}=b^{-\sum _{x}p(x)\log _{b}p(x)}} where x ranges over the events , where 0 −0 is defined to be 1 , and where the value of b does not affect the result; b can be chosen to be 2 , 10 , e , or any other positive value other than 1 . In some contexts, this measure is also referred to as the (order-1 true) diversity .

The logarithm log PP (p) is the entropy of the distribution; it is expressed in bits if the base of the logarithm is 2, and it is expressed in nats if the natural logarithm is used.

Perplexity of a random variable X may be defined as the perplexity of the distribution over its possible values x . It can be thought of as a measure of uncertainty or "surprise" related to the outcomes.

For a probability distribution p where exactly k outcomes each have a probability of 1 / k and all other outcomes have a probability of zero, the perplexity of this distribution is simply k . This is because the distribution models a fair k -sided die , with each of the k outcomes being equally likely. In this context, the perplexity k indicates that there is as much uncertainty as there would be when rolling a fair k -sided die. Even if a random variable has more than k possible outcomes, the perplexity will still be k if the distribution is uniform over k outcomes and zero for the rest. Thus, a random variable with a perplexity of k can be described as being " k -ways perplexed," meaning it has the same level of uncertainty as a fair k -sided die.

Perplexity is sometimes used as a measure of the difficulty of a prediction problem. It is, however, generally not a straightforward representation of the relevant probability. For example, if you have two choices, one with probability 0.9, your chances of a correct guess using the optimal strategy are 90 percent. Yet, the perplexity is 2 −0.9 log 2 0.9 - 0.1 log 2 0.1 = 1.38. The inverse of the perplexity, 1/1.38 = 0.72, does not correspond to the 0.9 probability.

The perplexity is the exponentiation of the entropy, a more commonly encountered quantity. Entropy measures the expected or "average" number of bits required to encode the outcome of the random variable using an optimal variable-length code . It can also be regarded as the expected information gain from learning the outcome of the random variable, providing insight into the uncertainty and complexity of the underlying probability distribution.

Perplexity of a probability model

A model of an unknown probability distribution p may be proposed based on a training sample that was drawn from p . Given a proposed probability model q , one may evaluate q by asking how well it predicts a separate test sample x 1 , x 2 , ..., x N also drawn from p . The perplexity of the model q is defined as

b − 1 N ∑ i = 1 N log b ■ q (x i) = (∏ i q (x i)) − 1 / N {\displaystyle b^{-{\frac {1}{N}}\sum _{i=1}^{N}\log _{b}q(x_{i})}=\left(\prod _{i}q(x_{i})\right)^{-1/N}}

where b {\displaystyle b} is customarily 2. Better models q of the unknown distribution p will tend to assign higher probabilities q (x i) to the test events. Thus, they have lower perplexity because they are less surprised by the test sample. This is equivalent to saying that better models have higher likelihoods for the test data, which leads to a lower perplexity value.

The exponent above may be regarded as the average number of bits needed to represent a test event x i if one uses an optimal code based on q . Low-perplexity models do a better job of compressing the test sample, requiring few bits per test element on average because q (x i) tends to be high.

The exponent − 1 N ∑ i = 1 N log b ■ q (x i) {\textstyle -{\frac {1}{N}}\sum _{i=1}^{N}\log _{b}q(x_{i})} may also be interpreted as a cross-entropy :

H (p ~ , q) = − ∑ x p ~ (x) log b ■ q (x) {\displaystyle H({\tilde {p}},q)=-\sum _{x}{\tilde {p}}(x)\log _{b}q(x)}

where p ~ {\displaystyle {\tilde {p}}} denotes the empirical distribution of the test sample (i.e., p ~ (x) = n / N {\displaystyle {\tilde {p}}(x)=n/N} if x appeared n times in the test sample of size N).

By the definition of KL divergence , it is also equal to H (p ~) + D K L (p ~ ■ q) {\displaystyle H({\tilde {p}})+D_{KL}({\tilde {p}}\|q)} , which is ≥ H (p ~) {\displaystyle \geq H({\tilde {p}})} . Consequently, the perplexity is minimized when q = p ~ {\displaystyle q={\tilde {p}}} .

Perplexity per token

In natural language processing (NLP), a corpus is a structured collection of texts or documents, and a language model is a probability distribution over entire texts or documents. Consequently, in NLP, the more commonly used measure is perplexity per token (word or, more frequently, sub-word), defined as: (∏ i = 1 n q (s i)) − 1 / N {\displaystyle \left(\prod _{i=1}^{n}q(s_{i})\right)^{-1/N}} where s 1 , . . . , s n {\displaystyle s_{1},...,s_{n}} are the n {\displaystyle n} documents in the corpus and N {\displaystyle N} is the number of tokens in the corpus. This normalizes the perplexity by the length of the text, allowing for more meaningful comparisons between different texts or models rather than documents.

Suppose the average text x i in the corpus has a probability of 2 − 190 {\displaystyle 2^{-190}} according to the language model. This would give a model perplexity of 2 190 per sentence. However, in NLP, it is more common to normalize by the length of a text. Thus, if the test sample has a length of 1,000 tokens, and could be coded using 7.95 bits per token, one could report a model perplexity of 2 7.95 = 247 per token. In other words, the model is as confused on test data as if it had to choose uniformly and independently among 247 possibilities for each token.

There are two standard evaluation metrics for language models: perplexity or word error rate (WER). The simpler of these measures, WER, is simply the percentage of erroneously recognized words E (deletions, insertions, substitutions) to total number of words N, in a speech recognition task i.e. W E R = (E N) × 100 % {\displaystyle {\mathit {WER}}=\left({\frac {E}{\mathrm {N} }}\right)\times 100\%} The second metric, perplexity (per token), is an information theoretic measure that evaluates the similarity of proposed model m to the original distribution p . It can be computed as a inverse of (geometric) average probability of test set T

P P L (D) = 1 m (T) N = 2 − 1 N log 2 ■ (m (T)) {\displaystyle {\mathit {PPL}}(D)={\sqrt[{N}]{\frac {1}{m(T)}}}=2^{-{\frac {1}{N}}\log _{2}\left(m(T)\right)}}

where N is the number of tokens in test set T . This equation can be seen as the exponentiated cross entropy, where cross entropy H (p ; m) is approximated as

H (p ; m) = − 1 N log 2 ■ (m (T)) {\displaystyle H(p;m)=-{\frac {1}{N}}\log _{2}\left(m(T)\right)}

Recent advances in language modeling

Since 2007, significant advancements in language modeling have emerged, particularly with the advent of deep learning techniques. Perplexity per token, a measure that quantifies the predictive

power of a language model, has remained central to evaluating models such as the dominant transformer models like Google's BERT , OpenAI's GPT-4 and other large language models (LLMs).

This measure was employed to compare different models on the same dataset and guide the optimization of hyperparameters , although it has been found sensitive to factors such as linguistic features and sentence length. [2]

Despite its pivotal role in language model development, perplexity has shown limitations, particularly as an inadequate predictor of speech recognition performance, overfitting and generalization , [3] [4] raising questions about the benefits of blindly optimizing perplexity alone.

Brown Corpus

The lowest perplexity that had been published on the Brown Corpus (1 million words of American English of varying topics and genres) as of 1992 is indeed about 247 per word/token, corresponding to a cross-entropy of log 2 247 = 7.95 bits per word or 1.75 bits per letter [5] using a trigram model. While this figure represented the state of the art (SOTA) at the time, advancements in techniques such as deep learning have led to significant improvements in perplexity on other benchmarks, such as the One Billion Word Benchmark. [6]

In the context of the Brown Corpus , simply guessing that the next word is "the" will achieve an accuracy of 7 percent, contrasting with the 1/247 = 0.4 percent that might be expected from a naive use of perplexity. This difference underscores the importance of the statistical model used and the nuanced nature of perplexity as a measure of predictiveness. [7] The guess is based on unigram statistics, not on the trigram statistics that yielded the perplexity of 247, and utilizing trigram statistics would further refine the prediction.

See also

Cross-entropy

Statistical model validation

References

v

t

e

MSE

MAE

sMAPE

MAPE

MASE

MSPE

RMS

RMSE/RMSD

R 2

MDA

MAD

F-score

P4

Accuracy

Precision

Recall

Kappa

MCC

AUC

ROC

Sensitivity and specificity

Logarithmic loss

Silhouette

Calinski–Harabasz index

Davies–Bouldin index

Dunn index

Hopkins statistic

Jaccard index

Rand index

Similarity measure

SMC

DBCV index

MRR

NDCG

AP

PSNR

SSIM

IoU

Perplexity

BLEU

Inception score

FID

Coverage

Intra-list similarity

Cosine similarity

Euclidean distance

Pearson correlation coefficient

Confusion matrix