Active Learning

参考：

[主动学习（Active Learning），看这一篇就够了 - 知乎](https://zhuanlan.zhihu.com/p/377045943)

[SupeRuier/awesome-active-learning: Everything you need about Active Learning (AL).](https://github.com/SupeRuier/awesome-active-learning?tab=readme-ov-file)

1. Active Learning in NLP

Reference: https://doi.org/10.48550/arXiv.2210.10109

* 1. Query Strategies
     1. Informativeness
        1. Uncertainty

1. Probalilistic
2. Entropy-based
3. Least confidence
4. Margin sampling
5. Check the divergence:

Nearest neighbor, adversarial perturbation, data augmentation.

* + - 1. Disagreement

Use multiple models and select the instances that are **most disagreed among them**.

1. Query-by-committee
2. Disagreement measured by:

Vote entropy, KL-divergence

1. Bayesian method
2. Dropout
   * + 1. Gradient

Choose the instances that would **most strongly impact the model**.

1. The strategy of expected gradient length (EGL)
2. Gradients of word embeddings
   * + 1. Performance prediction

The selected instances should be the ones that most **reduce future errors**. Computation costs a lot since retraining is needed for each candidate.

A similar and simpler idea is to select **the most erroneous or ambiguous instances** with regard to the current task model: train another smaller model to predict the losses.

* + 1. Representativeness

仅考虑informativeness会导致采样偏差和过多离群点。Representativeness考虑待选数据点之间的联系。

* + - 1. Density

In order to avoid the outliers, density-based strategies prefer instances that are **more representative of the unlabeled set**.

1. Measure an instance’s average similarity to all other instances.
2. K-nearest neighbor
   * + 1. Discriminative

Select instances that are **different from already labeled instances.**

1. Similarity scores
2. Train a model to discriminate the labeled and unlabeled sets.
   * + 1. Batch diversity

Select a batch of instances to train each time. （这个其实我们已经在使用了）

In this case, we need to consider the dissimilarities not only between selected instances and labeled ones but also **within the selected batch**.（这个我们没有考虑，**目标是选择一批diverse的样本**）

1. Iterative selection: in each iteration, an instance is selected by comparing it with previously chosen instances to avoid redundancy.
2. Clustering-based selection: part the unlabeled data into clusters and select instances among them.

To calculate similarity, in addition to comparing the input features or intermediate representations, other methods are also investigated, such as utilizing model-based similarity, gradients, and masked LM surprisal embeddings.

* + 1. Hybrid

Combine informativeness and representativeness.

1. Weighted sum / multiplication
2. Weighted clustering
3. Multi-step querying:

e.g. first high uncertain, then cluster to select a diverse batch from them.

1. Dynamic combination: different strategies may excel at different AL phases.