Advanced Active Learning Cheatsheet

Supervised Machine Learning models have two types of errors that can be fixed with more labeled data: errors that the models know about and errors that the models don’t yet know about. **Uncertainty Sampling** is the **Active Learning** strategy to find the known errors and **Diversity Sampling** is the strategy to find the unknown errors. This cheatsheet has 10 common methods to combine Uncertainty Sampling & Diversity Sampling. See my cheatsheets on each for background:  [http://bit.ly/uncertainty_sampling](http://bit.ly/uncertainty_sampling)  |  [http://bit.ly/diversity_sampling](http://bit.ly/diversity_sampling)

1. **Least Confidence Sampling with Clustering-based Sampling**: sample items that are confusing to your model and then cluster those items to ensure a diverse sample.

2. **Uncertainty Sampling with Model-based Outliers**: sample items that are confusing to your model and within those find items with low activation in the model.

3. **Uncertainty Sampling with Model-based Outliers and Clustering**: combine methods 1 and 2.

4. **Representative Cluster-based Sampling**: cluster your data to capture multinodal distributions and sample items that are most like your target domain.

5. **Sampling from the Highest Entropy Cluster**: cluster your unlabeled data and find the cluster with the highest average confusion for your model.

6. **Uncertainty Sampling and Representative Sampling**: sample items that are both confusing to your current model and the most like your target domain.

7. **Model-based Outliers and Representative Sampling**: sample items that have low activation in your model but are relatively common in your target domain.

8. **Clustering with itself for hierarchical clusters**: recursively cluster to maximize the diversity.

9. **Sampling from the Highest Entropy Cluster with Margin of Confidence Sampling**: find the cluster with the most confusion and then sample for the maximum pairwise label confusion within that cluster.

10. **Combining Ensemble Methods and Dropouts with individual strategies**: aggregate results that come from multiple models or multiple predictions from one model via Monte-Carlo Dropouts aka Bayesian Deep Learning.

**Tip: treat the individual Active Learning methods as building blocks to be combined:**
Uncertainty Sampling and Diversity Sampling work best in combination.
While academic papers about combining Uncertainty Sampling and Diversity Sampling focus on single metrics that combine the two, in practice you can simply chain the methods: apply one method to get a large sample and then refine that sample with another method.

See the book for more details on Active Learning building blocks and advanced methods for combining them, with open source implementations in PyTorch. [robertmunro.com](http://robertmunro.com) | [@VVRob](http://twitter.com/VVRob)