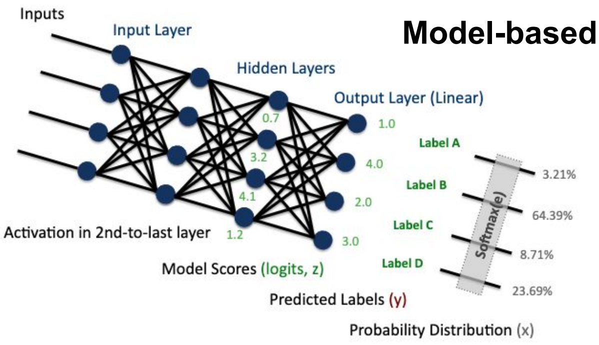
Diversity Sampling Cheatsheet

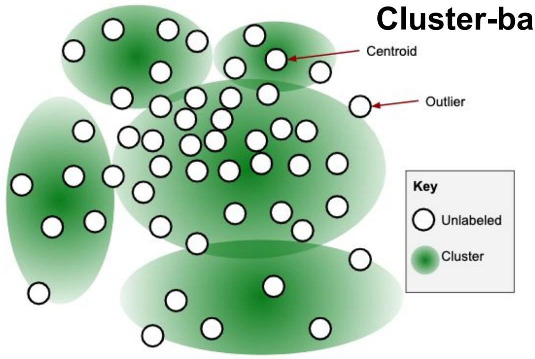
Supervised Machine Learning models are limited by their data. For example, a chat bot will not support diversity if trained only on one variety of English. For many tasks, you need to find data that represents diversity in the data and diversity in the real-world. This is a form of *Active Learning* known as *Diversity Sampling*.
This cheatsheet shares four ways to increase the diversity of your training data.



Model-based Outliers: sampling for low activation in logits and hidden layers

Why? To find items that are confusing to your model because of lack of information. This is different from uncertainty through *conflicting* information, a complementary sampling method.

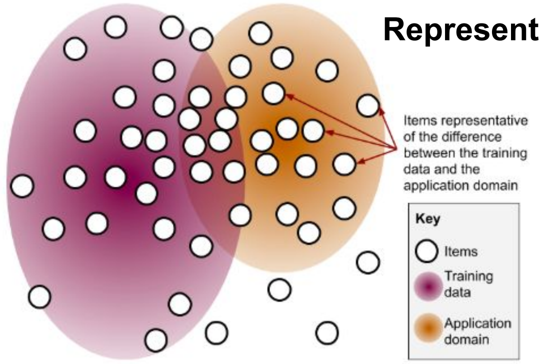
Tips: experiment with average vs max activation



Cluster-based Sampling: using unsupervised learning to pre-segment the data

Why? To ensure that you are sampling data from all the meaningful trends in your data's feature-space, not just the trends that contain the most items. Also to find outliers that are not part of any trend.

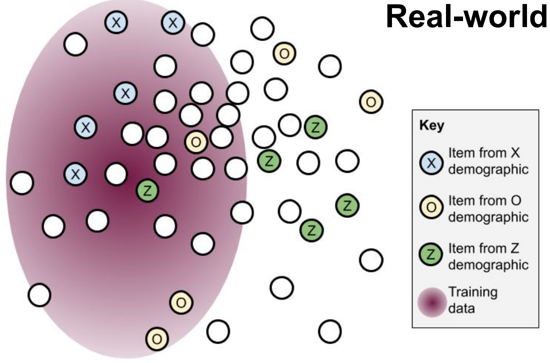
Tips: try different distance metrics and clustering algorithms


Items representative of the difference between the training data and the application domain
Key
Items
Training data
Application domain

Representative Sampling: finding items most representative of the target domain

Why? When your target domain is different from your current training data, you want to sample items *most* representative of your target domain in order to adapt to the domain as fast as possible.

Tips: extendable to be adaptive within one Active Learning cycle



Real-world diversity: increase fairness with data supporting real-world diversity

Why? So as many people can as possible take advantage of your models and you are not amplifying real-world biases. Use all Active Learning strategies to make your data as fair as possible.

Tips: your model might not require representative data to be fair

Robert Munro. **Human-in-the-Loop Machine Learning**, *Manning Publications*. http://bit.ly/huml_book
See the book for more details on each method and other Active Learning strategies like Uncertainty Sampling, with open source implementations in PyTorch. robertmunro.com | @WWRob