Active Learning for Computer Vision Curriculum

Project 2 Solution: Active Learning Introduction

Total Points Possible: 30

Problem 1 (15 points)

What is active learning and why is it important? How do you evaluate the performance of an active learning algorithm? What are the different categories of active learning?

In order to train a reliable machine learning model, a lot of labeled training data is needed. However, while gathering unlabeled data is cheap and easy, annotating the samples with class labels is an expensive process in terms of time and manual labor. Active learning algorithms automatically identify the salient and representative samples from vast amounts of unlabeled data. This tremendously reduces human annotation effort as only the unlabeled samples that are identified by the algorithm need to be labeled manually. Further, since the model gets trained on the salient and informative samples from the underlying population, its generalization capability is much better than a passive learner, where the training data is sampled at random. The importance of active learning lies in its capability of producing accurate learning models at minimal manual supervision.

Active learning can be broadly categorized as online and pool-based. In online active learning, the unlabeled samples are encountered sequentially over time. The active learner needs to decide on-the-fly whether a particular sample should be queried for its label. Samples discarded cannot be reused and the learner is not aware of the samples that are yet to arrive. In a pool-based active learning setup, the learner is exposed to a pool of unlabeled samples and it iteratively queries samples for annotation. Pool-based active learning is further classified as serial query based, where only a single unlabeled instance is queried in each iteration and batch mode, where a batch of unlabeled samples are queried simultaneously.

Problem 2 (15 points)

Briefly describe some of the challenges of active learning with examples. What is the difference between active and semi-supervised learning?

1. Noisy Oracles: A common assumption in active learning algorithms is that the human labeling oracles are flawless, that is, the annotations provided by the labeling oracles are always correct. This assumption may be violated under many practical situations. Consider a medical imaging application, where we are interested to develop a machine learning model to distinguish between

cancer and non-cancer images. We need a large number of labeled examples to train such a model. It is possible that some of the images may be incorrectly labeled by a junior resident (not having the expertise of an experience doctor) in this application. This necessitates the development of machine learning algorithms which considers the expertise and the reliability of labeling oracles in algorithm development.

2. Variable Labeling Costs: The cost of labeling each unlabeled example is assumed to be a constant in active learning applications. This may not be true always. Consider a voicemail classification problem where we are trying to classify voicemails as urgent or non-urgent. To annotate a voicemail in such an application, a human oracle has to listen to the entire voicemail before providing a label. Longer voicemails require more time to be labeled than shorter ones. This is an example of an application where the unlabeled samples have variable labeling costs, which needs to be factored in while developing the active learning algorithm.

In semi-supervised learning, the learner is exposed to both labeled and unlabeled training data. A model is learnt from both the labeled and unlabeled samples. There is no human effort involved. In contrast, in an active learning setup, the learner is exposed to a pool of unlabeled instances together with a budget. The learner can purchase the labels (get them annotated by a human expert) of some of the unlabeled samples. The goal is to design efficient query strategies to select the most informative unlabeled samples for manual labeling, so as to obtain the best machine learning model at minimal manual effort.