Data classification through the consensus of an ensemble

Jeremy Wong
University of Cambridge
What is machine learning and data classification

Ensemble methods

Student-teacher training
What is machine learning?

Machine learning

Regression

Classification

\[ y \]

\[ x \]

panda

car
Classification

* How do we make a decision?

* Bayes’ decision rule:  \( \hat{c} = \arg \max_{c'} P(c = c' | x, \theta) \)

* This leads to minimum classification error.
How do we train a classifier?

* We know true class training data points

\[ P^*(c = c^* | x) = \delta(c, c^*) \]

* Minimise a cost function

Kullback-Leibler divergence

\[ L(\theta) = \sum_i \sum_c P^*(c | x_i) \log \frac{P^*(c | x_i)}{P(c | x_i, \theta)} \]
Find $\theta^* = \arg\min_{\theta} L(\theta)$

$$= \arg\min_{\theta} \left\{ \sum_i \sum_c P^*(c|x_i) \log \frac{P^*(c|x_i)}{P(c|x_i, \theta)} \right\}$$

Ideally, want to solve $\frac{\partial L}{\partial \theta} = 0$

Use gradient descent $\theta^{(\tau)} = \theta^{(\tau-1)} - \eta \frac{\partial L}{\partial \theta}\bigg|_{\theta^{(\tau-1)}}$
Problem of using single model

- Multiple equivalent models

- True model may be outside space of possible models
Advantage of combining an ensemble

- Reduce the variance of the model estimate.
- Reduce overfitting and generalise better to unseen data.
- Correct each other’s errors.
- Expand the space of possible models.
How to combine multiple classifiers

* Mixture model / Bayesian average

\[ P(c|x, D) = \int P(c|x, \theta)p(\theta|D)d\theta \]

\[ \approx \sum_{m=1}^{M} P(c|x, \theta_m)p(\theta_m|D) \]

\[ \approx \frac{1}{M} \sum_{m=1}^{M} P(c|x, \theta_m) \]
Other ensemble methods

* Product of experts

\[
P(c|x, D) \approx \frac{1}{Z} \prod_{m=1}^{M} \left[ P(c|x, \theta_m) \right]^{\frac{1}{m}}
\]

* Majority voting

\[
P(c|x, D) \approx \frac{1}{M} \sum_{m=1}^{M} P'(c|x, \theta_m)
\]

where

\[
P'(c|x, \theta_m) = \delta \left( c, \arg \max_{c'} P(c'|x, \theta_m) \right)
\]
Problem with using an ensemble

- Slow, $M \times$ computational requirements
Can we teach a single model to emulate the performance of an ensemble?
What can the teacher provide in addition to the training data?
- Relative likelihoods between classes
- How to generalise to new data

Emulate the class probability of the teacher model

\[ P(c|x_i, \theta) \rightarrow P(c|x_i, \varphi) \]

Cost function

\[
L(\theta) = \sum_i \sum_c P(c|x_i, \varphi) \log \frac{P(c|x_i, \varphi)}{P(c|x_i, \theta)}
\]
Experimental Results

* Speech recognition task trained on 3 hrs of data in the Tok Pisin language

<table>
<thead>
<tr>
<th>Model</th>
<th>WER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single model average</td>
<td>48.2</td>
</tr>
<tr>
<td>Teacher ensemble</td>
<td>47.0</td>
</tr>
<tr>
<td>Student</td>
<td>47.3</td>
</tr>
</tbody>
</table>
Thank you