– IN5550 –

Neural Methods in Natural Language Processing

Ensembles, transfer and multi-task learning: Part 1

Erik Velldal

Language Technology Group (LTG)
University of Oslo

– IN5550 –
Neural Methods in Natural Language Processing

Ensembles, transfer and multi-task learning: Part 1

Erik Velldal

Language Technology Group (LTG)
University of Oslo

This session

- No new bricks.
- Taking what we already have, putting it together in new ways.

This session

- No new bricks.
- Taking what we already have, putting it together in new ways.

- Ensemble learning
- Multi-task learning
- Transfer learning

This session

- No new bricks.
- Taking what we already have, putting it together in new ways.

- Ensemble learning
  - Training several models to do one task.
- Multi-task learning
- Transfer learning

This session

- No new bricks.
- Taking what we already have, putting it together in new ways.

- Ensemble learning
  - Training several models to do one task.
- Multi-task learning
  - Training one model to do several tasks.
- Transfer learning

This session

- No new bricks.
- Taking what we already have, putting it together in new ways.

- Ensemble learning
  - Training several models to do one task.
- Multi-task learning
  - Training one model to do several tasks.
- Transfer learning
  - Training a model for a new task based on a model for some other task.

Standard approach to model selection

- Train a bunch of models
- Keep the model with best performance on the development set
- Discard the rest

Standard approach to model selection

- Train a bunch of models
- Keep the model with best performance on the development set
- Discard the rest

- Some issues:
- Best on dev. is not necessarily best on held-out.
- ANNs generally have low bias and high variance, can be unstable and have a danger of overfitting.
- Models might have non-overlapping errors.
- Ensemble methods may help.

Ensemble learning

- Combine multiple models to obtain better performance than for any of the individual base models alone.
- The various base models in the ensemble could be based on the same or different learning algorithms.
- Several meta-heuristics available for how to create the base models and how to combine their predictions.

Ensemble learning

- Combine multiple models to obtain better performance than for any of the individual base models alone.
- The various base models in the ensemble could be based on the same or different learning algorithms.
- Several meta-heuristics available for how to create the base models and how to combine their predictions. E.g.:
  - Boosting
  - Bagging
  - Stacking
  - Mixture of Experts

Examples of ensembling

Boosting

- The base learners are generated sequentially:
- Incrementally build the ensemble by training each new model to emphasize training instances that previous models misclassified.
- Combine predictions through a weighted majority vote (classification) or average (regression).

Examples of ensembling

Boosting

- The base learners are generated sequentially:
- Incrementally build the ensemble by training each new model to emphasize training instances that previous models misclassified.
- Combine predictions through a weighted majority vote (classification) or average (regression).

Bagging (Bootstrap AGGregating)

- The base learners are generated independently:
- Create multiple instances of the training data by sampling with replacement, training a separate model for each.
- Combine ('aggregate') predictions by voting or averaging.

Examples of ensembling

Stacking

- Train several base-level models on the complete training set,
- then train a meta-model with the base model predictions as features.
- Often used with heterogeneous ensembles.

Examples of ensembling

Stacking

- Train several base-level models on the complete training set,
- then train a meta-model with the base model predictions as features.
- Often used with heterogeneous ensembles.

Mixture of Experts

- A meta-model approach like stacking.
- Use the input data to decide which model to rely on for each prediction:
- Determined by a gating network ('the manager') trained together with the base networks ('the experts').

Drawbacks of ensembling

- ANNs often applied in ensembles to squeeze out some extra F1 points.
- But their high leaderboard ranks come at a high computational cost:
- Must learn, store, and apply several separate models.
- Often not practical for deployment.

Distillation

- High acc./F1 models tend to have a high number of parameters.
- Often too inefficient to deploy in real systems.
- Knowledge distillation is a technique for reducing the complexity while retaining much of the performance.
- Idea: Train a (smaller) student model to mimic the behaviour of a (larger) teacher model.
- The student is typically trained using the output probabilities of the teacher as soft labels.
- Can be used to distill an ensemble into a single model.

Next

- Multi-task learning