Deep Density Networks with Uncertainty for spontaneous spoken language assessment

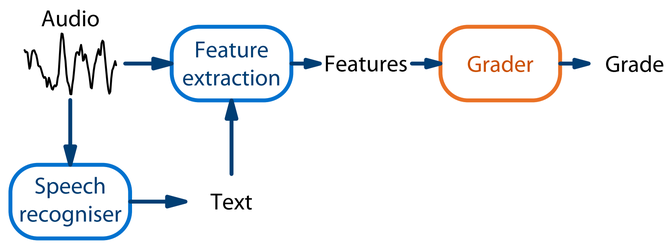
Andrey Malinin, Yu Wang, Kate Knill and Mark Gales
{am969,yw396,kate.knill,mjfg}@eng.cam.ac.uk

ALTA Institute / Department of Engineering, University of Cambridge

UNIVERSITY OF CAMBRIDGE

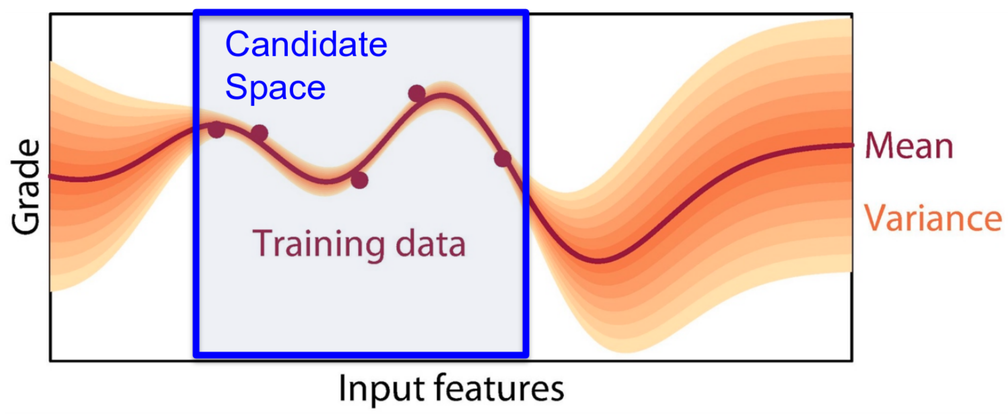
Introduction

- Many people are learning English → want official qualifications
- To help meet this demand: Automatic assessment of spoken English



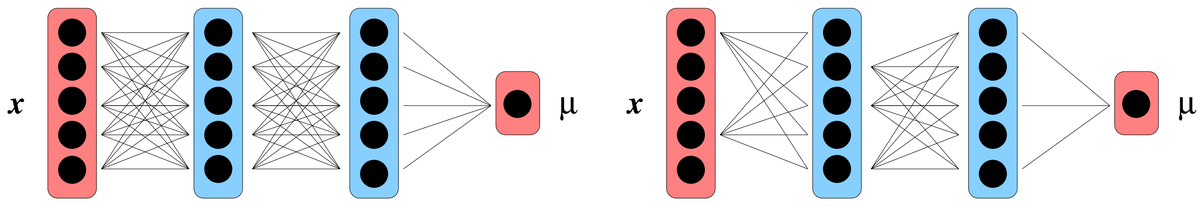
- An automatic grader:
  - is more consistent than human graders
  - has significantly higher throughput
- How to deal with difficult to grade speakers?
- Estimate uncertainty in prediction →
  - Reject speakers with greatest uncertainty to human graders

Uncertainty in Gaussian Processes



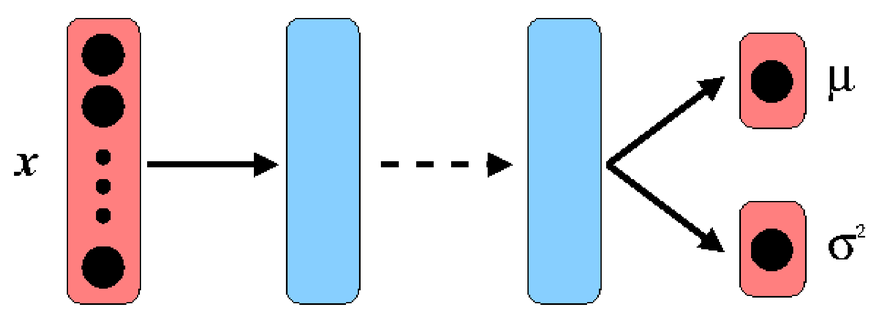
- Powerful non-parametric Bayesian model:
  - $\phi_{GP}(\mathbf{x}|\mathcal{D}) \rightarrow \mu_\phi(\mathbf{x}),\ \sigma^2_\phi(\mathbf{x})$
- Produces a heteroscedastic distribution over grades
- Uncertainty depends on proximity of test data to training data
- Limitations - $O(n^2)$ memory usage $O(n^3)$ compute load

Uncertainty in Deep Neural Networks



- Parametric model: $\phi_{DNN}(\mathbf{x}|\boldsymbol{\theta}) \rightarrow \mu_\phi(\mathbf{x})$
- Advantages - scalable and flexible architecture
- Limitation - No natural uncertainty measure
- Can derive an approximate bayesian measure via:
  - Monte-Carlo Dropout
  - Prediction uncertainty depends on uncertainty in weights

Deep Density Network



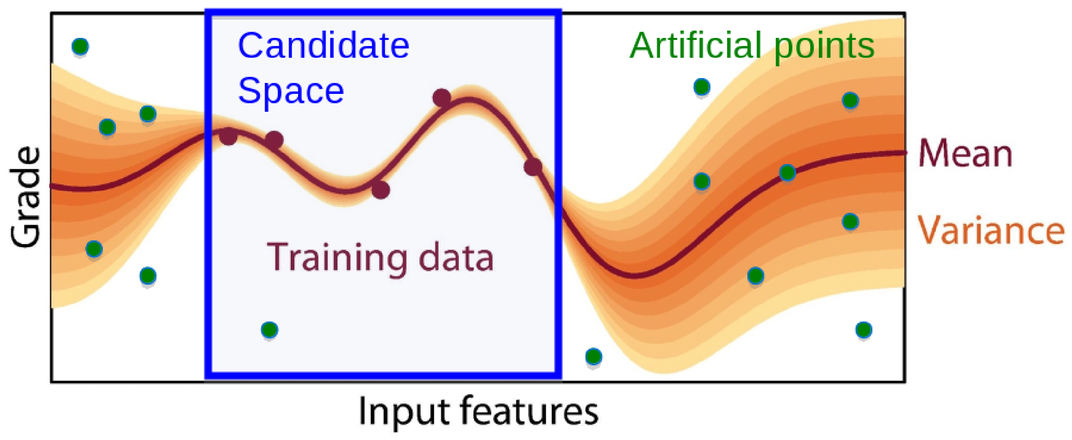
- DDN parametrises a normal distribution $\mathtt{Q}(\mathrm{y}|\mathbf{x};\boldsymbol{\theta})$ over grades.
  - Produces a heteroscedastic distribution.

$$\phi_{DDN}(\mathbf{x}|\boldsymbol{\theta}) \rightarrow \mu_\phi(\mathbf{x}),\ \sigma^2_\phi(\mathbf{x})$$
$$\mathtt{Q}(\mathrm{y}|\mathbf{x};\boldsymbol{\theta}) = \mathcal{N}(y|\mu_\phi(\mathbf{x}), \sigma^2_\phi(\mathbf{x}))$$

- Train by maximizing likelihood
- DDN variance represents the spread in grade given input $\mathbf{x}$ →
  - Natural noise associated with the data
- Desire to emulate GP uncertainty →
  - Assign uncertainty based on similarity to training data

Deep Density Network with Noise

- Need variance to depend on distance of $\mathbf{x}$ from training data
  - Low/High variance near/far from training data
- Solution: Specify variance directly
  - Define a low variance empirical distribution $\mathtt{P_D}$ over real data
  - Define a high-variance artificial data distribution $\mathtt{P_N}$
  - Train DDN to model both distributions



- Two stage training process:
  1. Train standard DDN on real data
  2. Continue training DDN in multi-task fashion →
     - Minimize KL divergence of $\mathtt{Q}(\mathrm{y}|\mathbf{x};\theta)$ to $\mathtt{P_D}$ and $\mathtt{P_N}$

$$\mathcal{L}(\boldsymbol{\theta}) = \mathcal{L}_D(\boldsymbol{\theta}) + \alpha\mathcal{L}_N(\boldsymbol{\theta})$$
$$= \mathtt{E}_{\hat{\mathbf{x}}}[\mathtt{KL}(\mathtt{P_D}(\mathrm{y}|\hat{\mathbf{x}})||\mathtt{Q}(\mathrm{y}|\hat{\mathbf{x}}))] + \alpha\mathtt{E}_{\tilde{\mathbf{x}}}[\mathtt{KL}(\mathtt{P_N}(\mathrm{y}|\tilde{\mathbf{x}})||\mathtt{Q}(\mathrm{y}|\tilde{\mathbf{x}}))]$$

Evaluation Metrics, Data and Experiments

- Grader Performance Assessment:
  - Pearson Correlation Coefficient (PCC)
  - Mean Squared Error (MSE)
- Useful to have a single value to represent rejection performance
  - Assess using Area Under Curve Rejection Ratio $\mathtt{AUC_{RR}}$

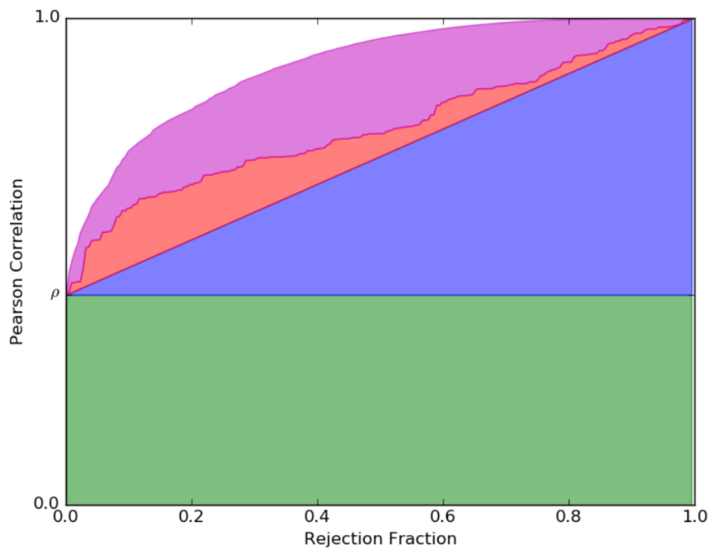
$$AUC_{RR} = \frac{AUC_{var}}{AUC_{max}}$$

■ $= AUC_{var}$

■ + ■ $= AUC_{max}$

$\mathtt{AUC_{RR}}$ = 0 = random rejection

$\mathtt{AUC_{RR}}$ = 1 = optimal rejection



Experiments

- Acoustic and ASR-derived features from spontaneous responses
- 29 dim. input features, single grade (0-30) target
- 4300 training and 230 evaluation speakers
- Training on standard grades
- Evaluation in expert grades

Grader	PCC	MSE	$\mathtt{AUC_{RR}}$
GP	0.857	9.8	0.234
DNN+MCDC	0.865	9.0	0.088
DDN	0.851	10.2	0.107
DDN+Noise	0.851	10.2	0.356

Conclusions

- DDN based approach:
  - Has comparable grading performance as GP and DNN
  - Provides a better uncertainty measure for rejection.
  - Combines essence of GP uncertainty with scalability of DNN
- Rejection performance metric $\mathtt{AUC_{RR}}$ developed
- Future Work:
  - $\mathtt{P_N}$ is a crude approximation → Factor Analysis, Variational Autoencoders