Uncertainty modelling in spoken language assessment

Dr. Jeremy Wong

Aural and Language Intelligence department

Institute for Infocomm Research (I^{2}R)

2nd October 2023

ARES PUBLIC

CREATING GROWTH, ENHANCING LIVES

Content

About myself

Background:

- Spoken language assessment
- Uncertainty estimation
- Neural network
- Gaussian process

Recent developments:

- Learning data uncertainty in a neural network
- Learning data uncertainty in a Gaussian process
- Learning distributional uncertainty from a Gaussian process
- Improving model assumptions

About myself

PhD in University of Cambridge, UK – 2014 to 2019

Research topic: speech recognition

Senior applied scientist in Microsoft, USA – 2019 to 2021

Research topic: speaker diarisation

Senior scientist in I^{2}R A*STAR, Singapore – 2021 to now

Research topic: spoken language assessment

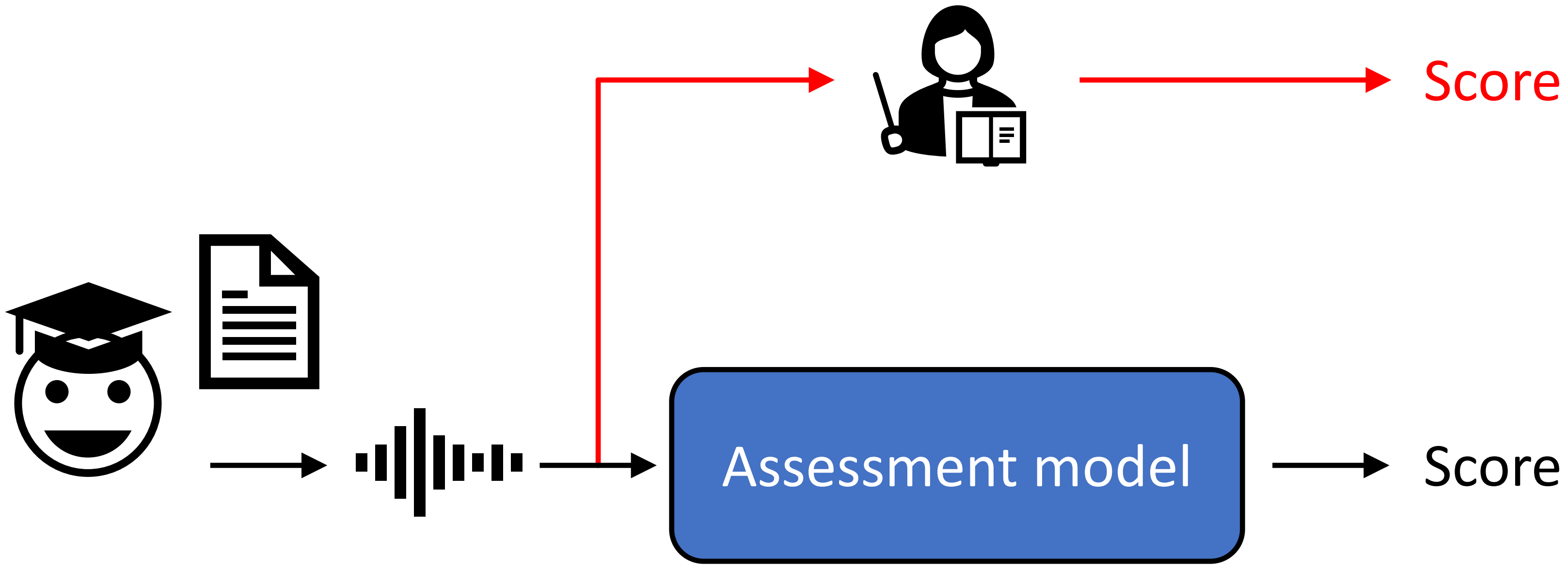
Spoken language assessment

MEMORIES FOR THE FUTURE
2002 - 2022

ARES PUBLIC

Institute for
Infocomm Research
I²R

Spoken language assessment



Spoken language assessment

Aspects to assess:

- Pronunciation accuracy
- Fluency, intonation, prosody
- Sentence completion
- Task completion
- Topic relevance

Applications:

- Automatic language tutoring
- Language practice
- Language examination

Dataset

<u>Speechocean762</u>

Training set: 2500 sentence, 125 speakers
Test set: 2500 sentences, 125 speakers

Annotation levels: sentence, word, phone
Annotation types:

- Pronunciation accuracy
- Fluency
- Prosody
- Sentence completion
- Word stress

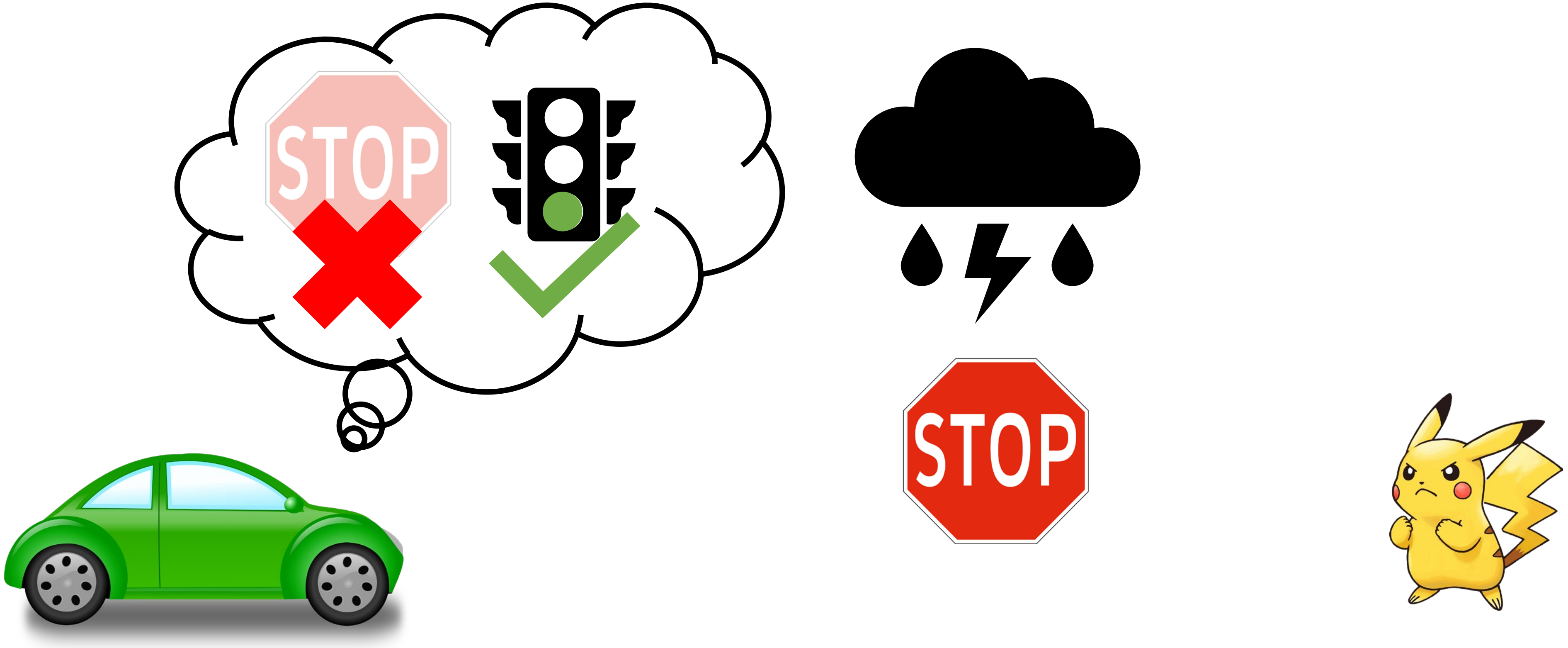
Evaluate model performance

Evaluation metrics:

- Pearson’s correlation coefficient
- Mean squared error

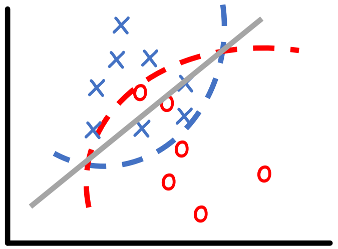
Uncertainty estimation

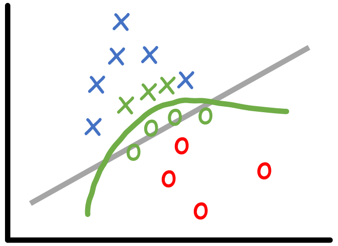
Uncertainty estimation

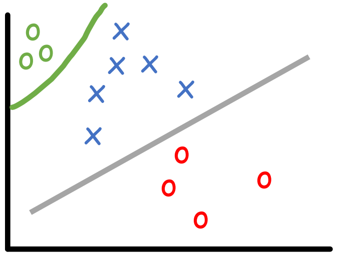


Types of uncertainty

	Data uncertainty	Model uncertainty	Distributional uncertainty
Caused by	Natural overlap in the input space. Limited access to information.	Each model architecture has an intrinsic bias toward certain behaviour. When given a finite training data, the optimal model architecture or parameter set is not unique. Multiple non-equivalent optima.	Finite coverage of the training data.
Alleviated by	Using more independent input features. (E.g. multi-modal) Using models that can process more aspects of the data. (E.g. RNN vs DNN)	Get more training data. Combine multiple models.	Increase the distributional support of the training data. (E.g. domain adaptation)







- How to get a model to know that it does not know?

ARES PUBLIC

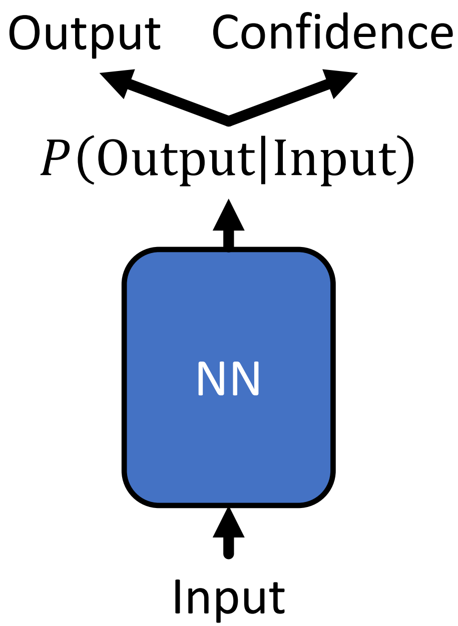
Use of uncertainty

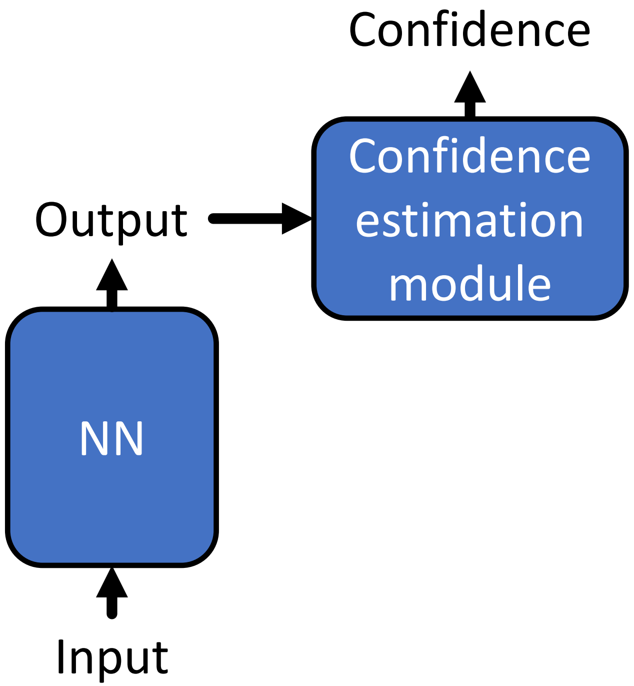
- Take precaution
  - Slow down
  - If multiple teachers would disagree, then don't penalise student

- Seek clarification from user
  - Ask the student to repeat or rephrase

- Seek human intervention
  - Ask a human teacher to assess the student instead

How to compute uncertainty





References about uncertainty

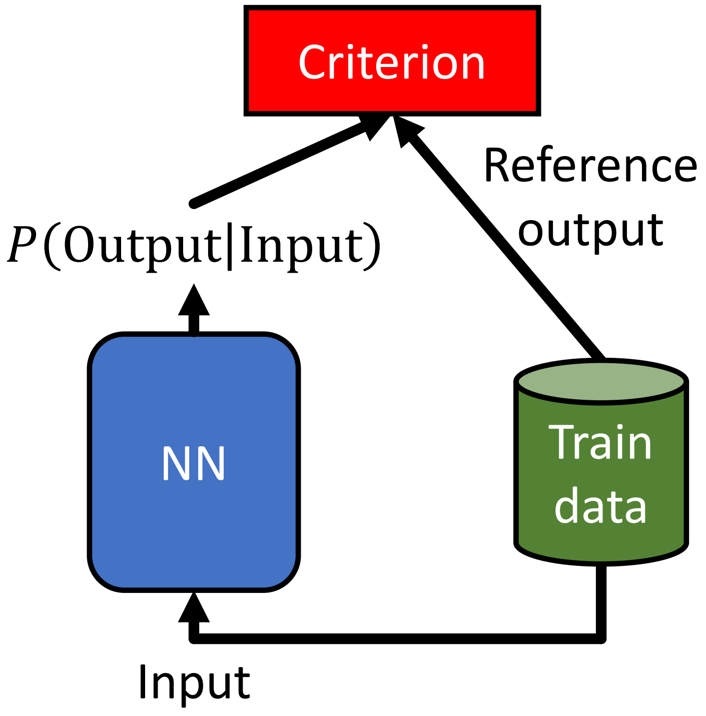
- A. Kendall and Y. Gal, *"What uncertainties do we need in Bayesian deep learning for computer vision?,"* NIPS, 2017
- A. Malinin and M. Gales, *"Predictive uncertainty estimation via prior networks,"* NeurIPS, 2018
- R. McAllister, Y. Gal, A. Kendall, M. van der Wilk, A. Shah, R. Cipolla, and A. Weller, *"Concrete problems for autonomous vehicle safety: advantages of Bayesian deep learning,"* IJCAI, 2017

Neural networks and Gaussian processes

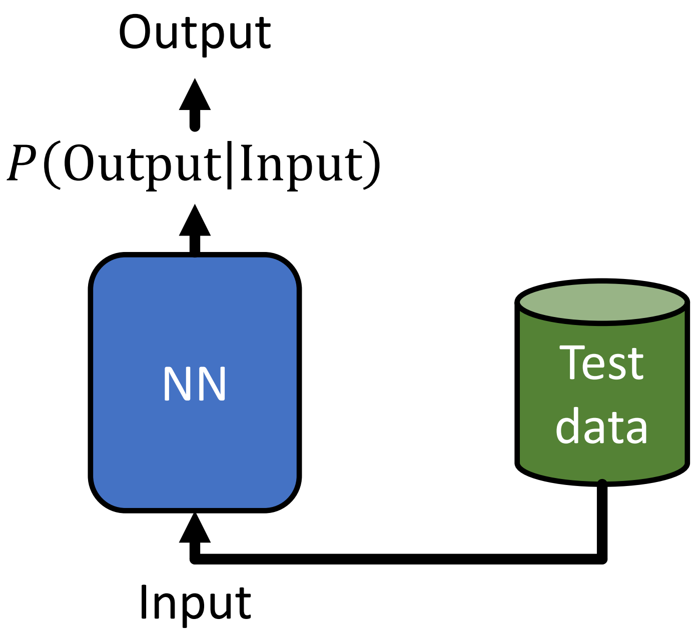
Institute for Infocomm Research
I²R

Neural network

Training

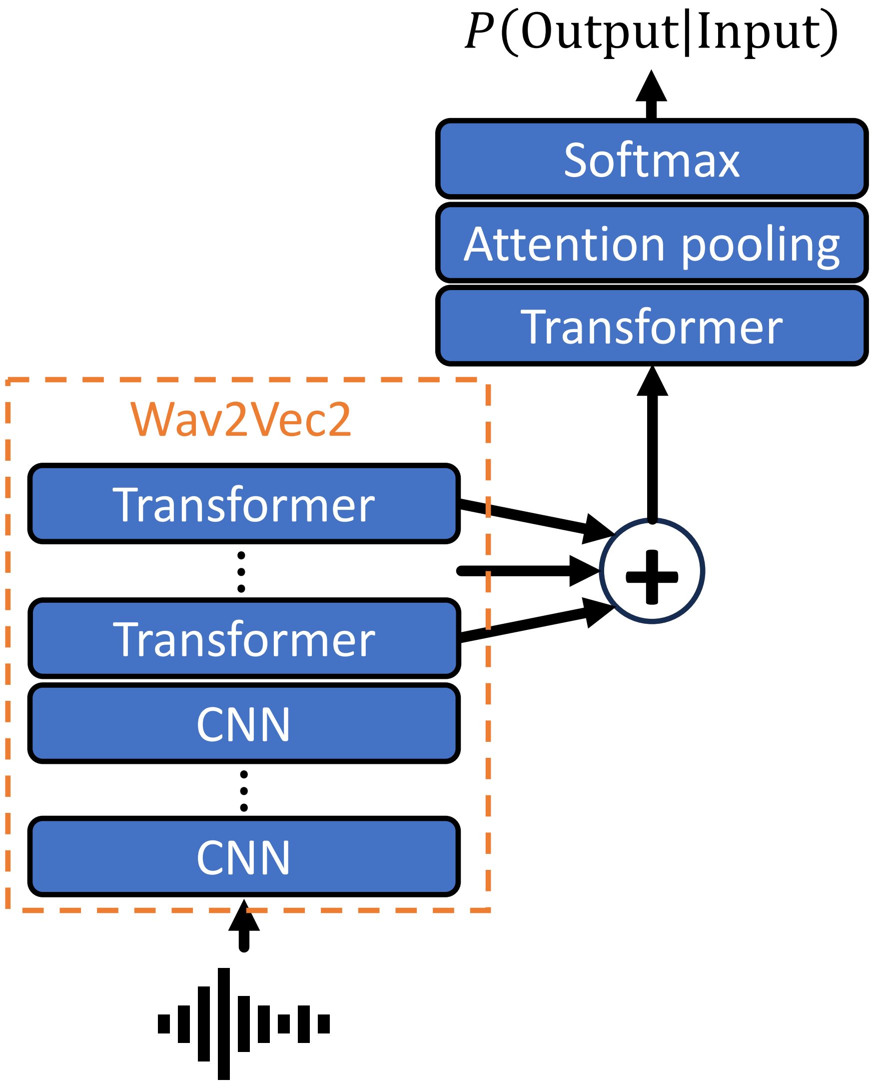


Inference



Example model for spoken language assessment

- Y. Gong, Z. Chen, I.-H. Chu, P. Chang, and J. Glass, *"Transformer-based multi-aspect multi-granularity non-native English speaker pronunciation assessment,"* ICASSP, 2022
- F.-A. Chao and T.-H. Lo and T.-I. Wu and Y.-T. Sung and B. Chen, *"3M: an effective multi-view, multi-granularity, and multi-aspect modeling approach to English pronunciation assessment,"* APSIPA, 2022
- S. Banno and M. Matassoni, *"Proficiency assessment of L2 spoken English using Wav2Vec 2.0,"* SLT, 2022



x_i -> input
y_i -> model output
y_i^{ref} -> reference output
θ -> model parameters

Training criteria

- Cross-entropy

$$\arg\max_{\theta} \sum_i \log P\left(y_i^{\text{ref}} \middle| x_i; \theta\right)$$

- Mean squared error

$$\arg\min_{\theta} \sum_i \left(y_i^{\text{ref}} - y_i\right)^2$$

x -> input
y -> output
$R(y, y')$-> risk

Inference decoding

- Maximum a-posteriori

$$\arg\max_{y} P(y|x)$$

- Mean

$$\sum_{y} yP(y|x)$$

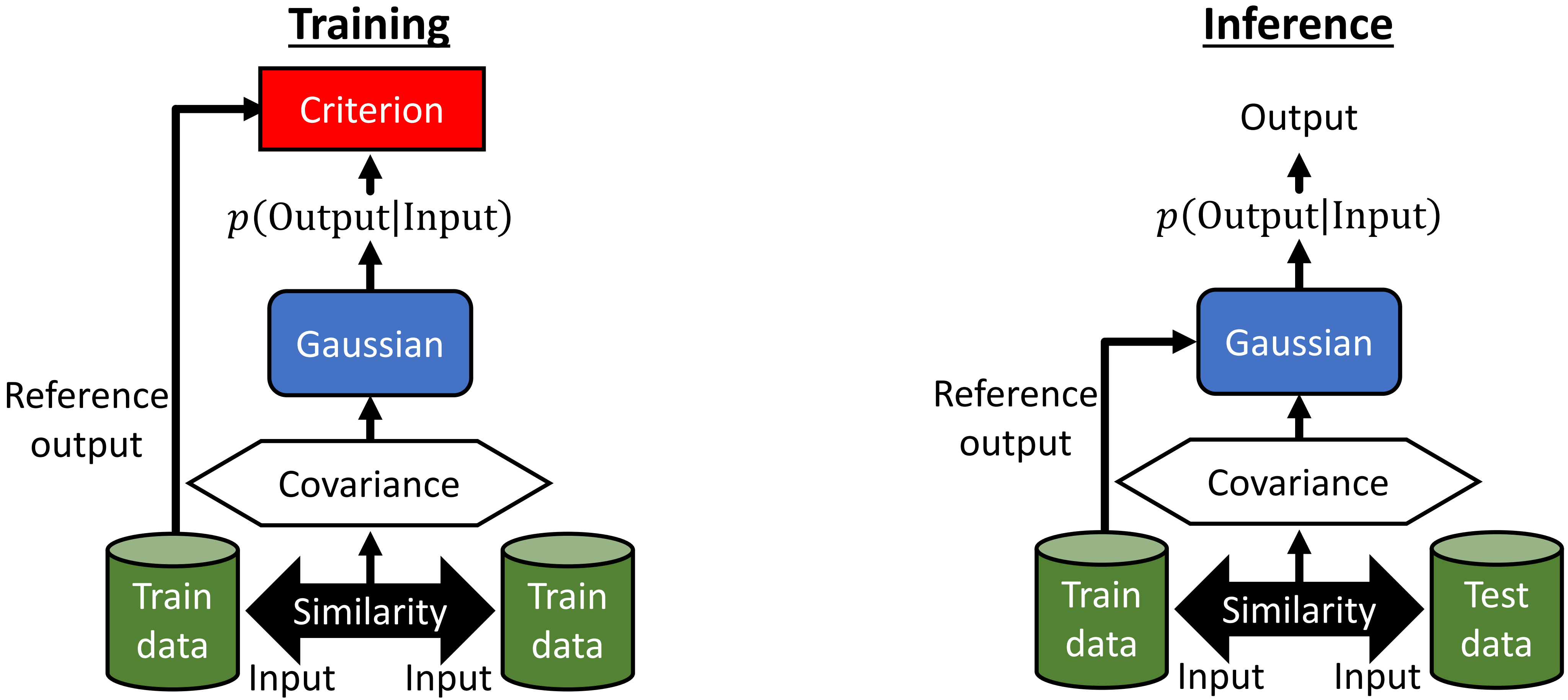
- Median

$$\arg\min_{y} y : \sum_{y'}^{y} P(y'|x) \geq \frac{1}{2}$$

- Minimum expected risk

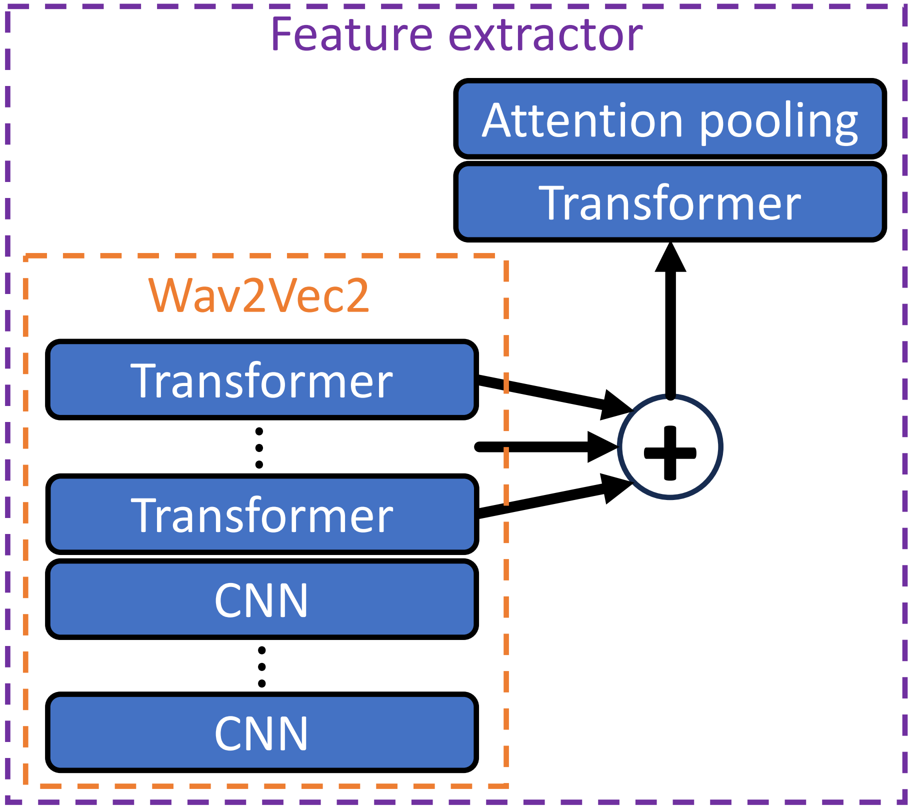
$$\arg\min_{y} \sum_{y'} R(y, y')P(y'|x)$$

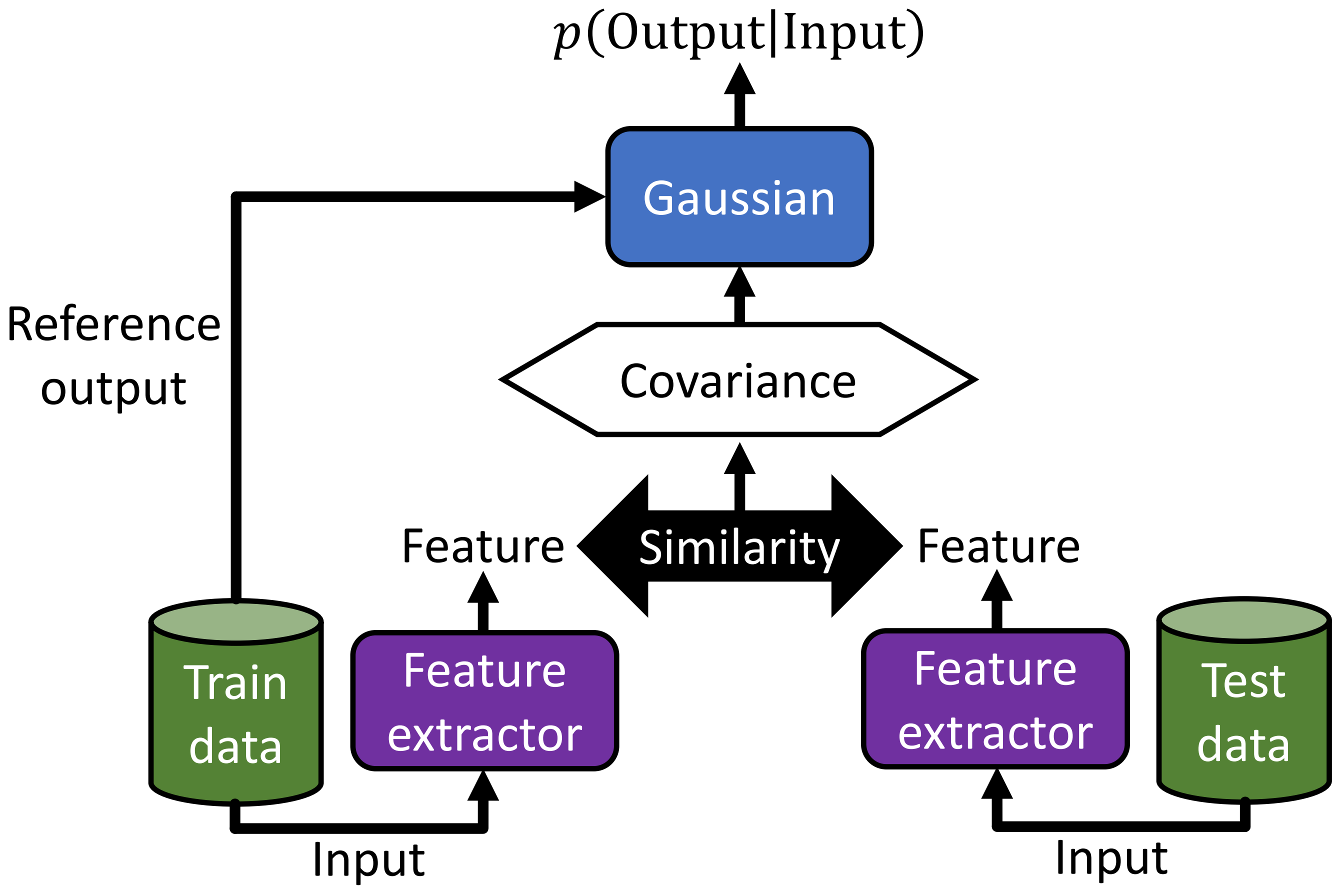
Gaussian process



C. Rasmussen and C. Williams, *"Gaussian processes for machine learning,"* MIT Press, 2006

Example model for spoken language assessment





Gaussian process formulation

$\boldsymbol{x}$ -> input
$\boldsymbol{y}$ -> model output
$\boldsymbol{f}$ -> latent variable
s, l, σ -> hyper-parameters
$\boldsymbol{I}$ -> Identity matrix
$\boldsymbol{0}$ -> Zero vector
$\mathcal{N}$ -> Gaussian

- Kernel

$$k_{ij}(\boldsymbol{x}, \boldsymbol{x}') = s^{\boldsymbol{2}} \exp\left[-\frac{\left(x_i - x_j'\right)^2}{2l^2}\right]$$

- Prior

$$p(\boldsymbol{f}|\boldsymbol{x}) = \mathcal{N}\left(\boldsymbol{f}; \boldsymbol{0}, \boldsymbol{K}(\boldsymbol{x}, \boldsymbol{x})\right)$$

- Output density function

$$p(\boldsymbol{y}|\boldsymbol{f}) = \mathcal{N}(\boldsymbol{y}; \boldsymbol{f}, \sigma^2 \boldsymbol{I})$$

- Marginal likelihood

$$p(\boldsymbol{y}|\boldsymbol{x}) = \int p(\boldsymbol{y}|\boldsymbol{f}) p(\boldsymbol{f}|\boldsymbol{x}) d\boldsymbol{f}$$
$$= \mathcal{N}(\boldsymbol{y}; \boldsymbol{0}, \boldsymbol{K}(\boldsymbol{x}, \boldsymbol{x}) + \sigma^2 \boldsymbol{I})$$

Gaussian process formulation

- Joint prior

$$p(\hat{f},\boldsymbol{y}|\hat{x},\boldsymbol{x}) = \mathcal{N}\left(\begin{bmatrix}\hat{f}\\ \boldsymbol{y}\end{bmatrix};\boldsymbol{0},\begin{bmatrix}k(\hat{x},\hat{x}) & \boldsymbol{k}^{\mathrm{T}}(\boldsymbol{x},\hat{x})\\ \boldsymbol{k}(\boldsymbol{x},\hat{x}) & \boldsymbol{K}(\boldsymbol{x},\boldsymbol{x})+\sigma^2\boldsymbol{I}\end{bmatrix}\right)$$

- Latent posterior

$$\begin{aligned} p(\hat{f}|\boldsymbol{y},\hat{x},\boldsymbol{x}) &= \frac{p(\hat{f},\boldsymbol{y}|\hat{x},\boldsymbol{x})}{p(\boldsymbol{y}|\boldsymbol{f})} \\ &= \mathcal{N}(\hat{f};\hat{\mu},\hat{v}) \end{aligned}$$

$$\begin{aligned} \hat{\mu} &= \boldsymbol{k}^{\mathrm{T}}(\boldsymbol{x},\hat{x})[\boldsymbol{K}(\boldsymbol{x},\boldsymbol{x})+\sigma^2\boldsymbol{I}]^{-\boldsymbol{1}}\boldsymbol{y} \\ \hat{v} &= k(\hat{x},\hat{x}) - \boldsymbol{k}^{\mathrm{T}}(\boldsymbol{x},\hat{x})[\boldsymbol{K}(\boldsymbol{x},\boldsymbol{x})+\sigma^2\boldsymbol{I}]^{-\boldsymbol{1}}\boldsymbol{k}(\boldsymbol{x},\hat{x}) \end{aligned}$$

- Output posterior

$$\begin{aligned} p(\hat{y}|\boldsymbol{y},\hat{x},\boldsymbol{x}) &= \int p(\hat{y}|\hat{f})p(\hat{f},\boldsymbol{y}|\hat{x},\boldsymbol{x})d\hat{f} \\ &= \mathcal{N}(\hat{y};\hat{\mu},\hat{v}+\sigma^2) \end{aligned}$$

Training and inference

Training criterion

- Maximum marginal log-likelihood

$$\arg\max_{\theta} \log p(\boldsymbol{y}|\boldsymbol{x})$$

Inference decoding

- For Gaussian, max = mean = median = $\hat{\mu}$.

Compare NN to GP

	NN	GP
Parameters	Many parameters to learn training data.	Only 3 hyper-parameters.
Inference	Training data not using during inference.	Training data used during inference. High computational cost.
Uncertainty	**Data:** to some extent **Distributional:** no **Model:** no	**Data:** no **Distributional:** yes **Model:** to some extent

Learning data uncertainty in a neural network

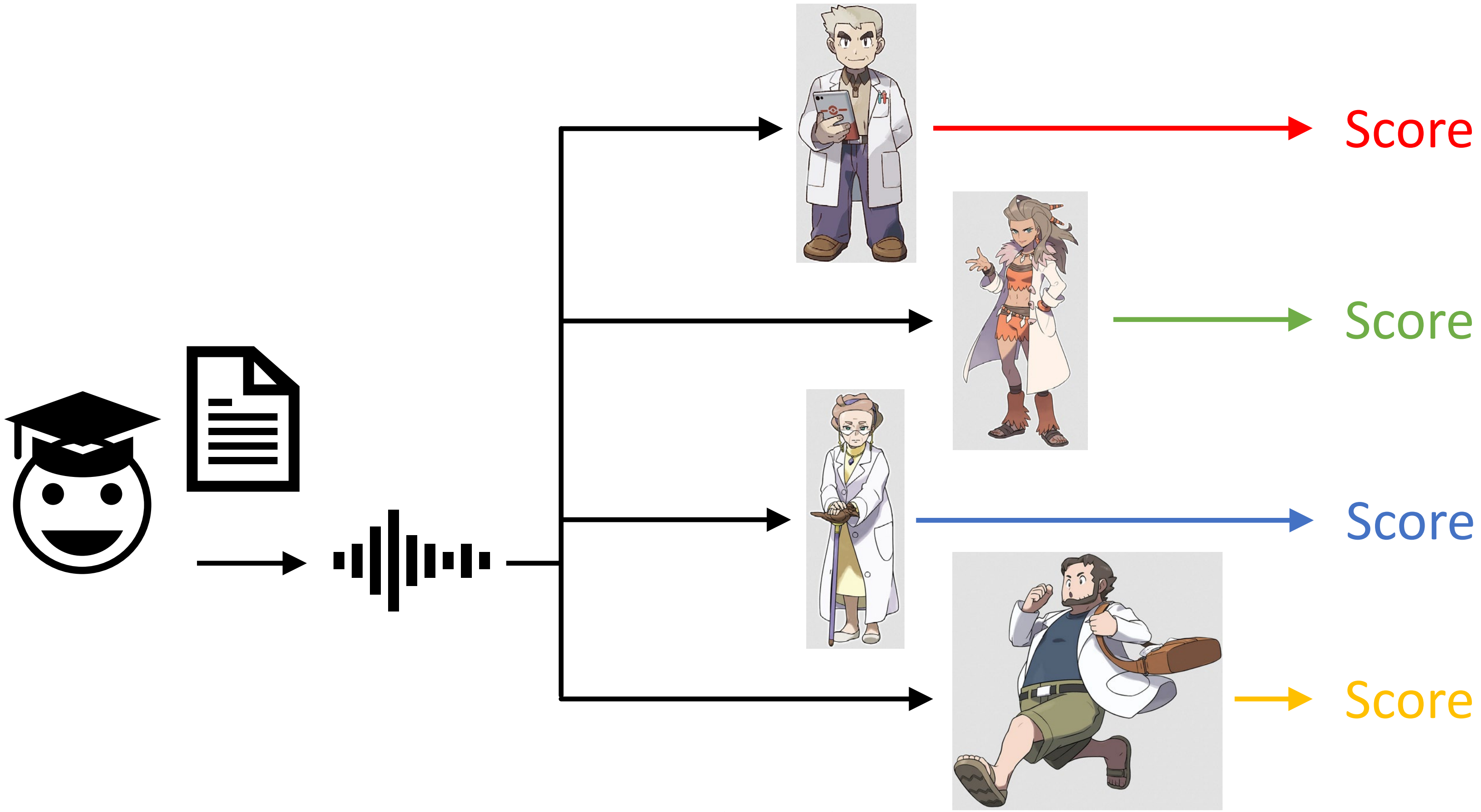
MEMORIES FOR THE FUTURE
2002 - 2022

ARES PUBLIC

Institute for Infocomm Research
I²R

Subjective data uncertainty

- Different human experts may not agree about what the correct output should be.

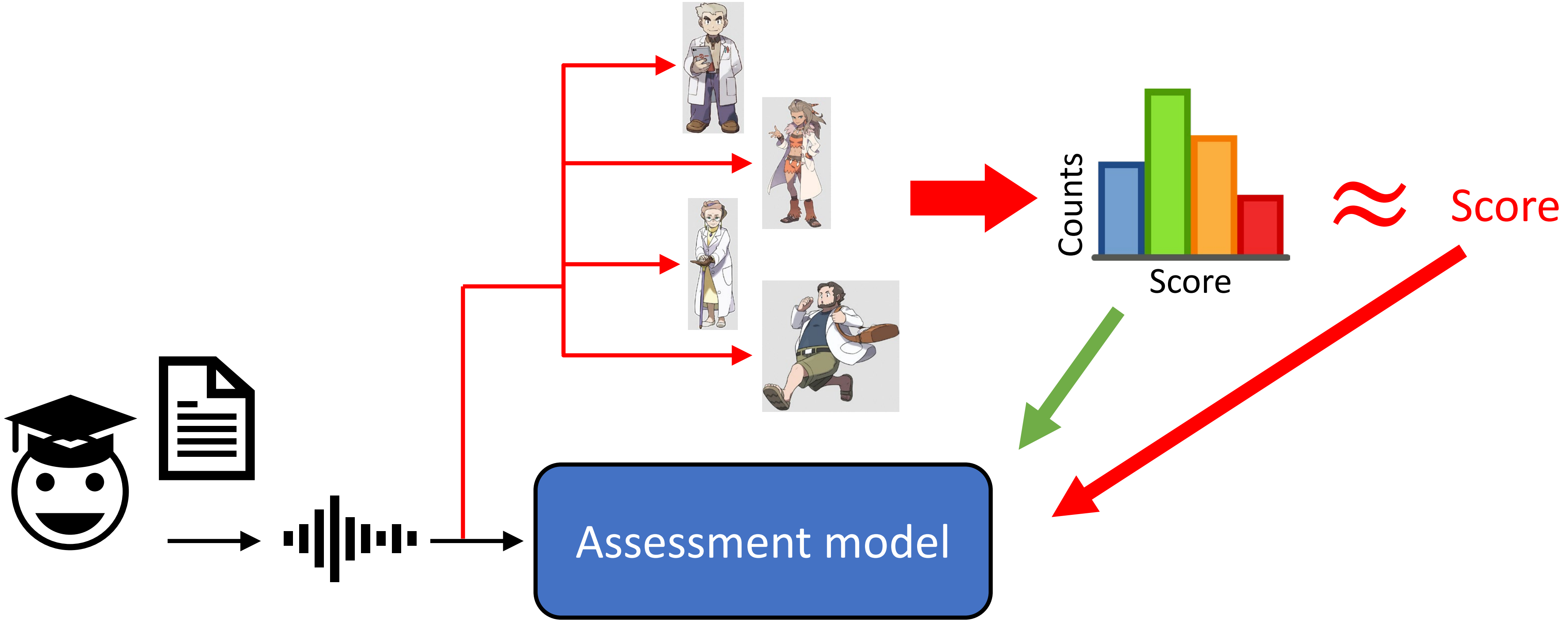


Importance of modelling data uncertainty

- If multiple teachers would disagree, then don't penalise student.

- Use uncertainty information to:
  - Ask the student to repeat or rephrase.
  - Ask a human teacher to assess the student instead.

Training to capture data uncertainty



- Collection of outputs from multiple humans forms reference of data uncertainty.
- Train and evaluate model using distance between reference and predicted distributions.

J. Wong, H. Zhang, and N. Chen, *"Modelling inter-rater uncertainty in spoken language assessment,"* IEEE Transactions on Audio, Speech, and Language Processing, vol. 31, Jul 2023

Learning data uncertainty in a Gaussian process

ARES PUBLIC

Institute for Infocomm Research

I²R

Using multiple reference outputs in GP

- Standard GP assumes each training input has 1 reference output.
- Extend GP to consider multiple training reference outputs.

- Training, joint marginal log-likelihood:

$$\arg\max_{\theta} \log p(\boldsymbol{y}_1, \cdots, \boldsymbol{y}_R | \boldsymbol{x})$$

- Inference, posterior:

$$p(\hat{y} | \boldsymbol{y}_1, \cdots, \boldsymbol{y}_R, \hat{x}, \boldsymbol{x})$$

J. Wong, H. Zhang, and N. Chen, *"Variational Gaussian process data uncertainty,"* ASRU, Dec 2023

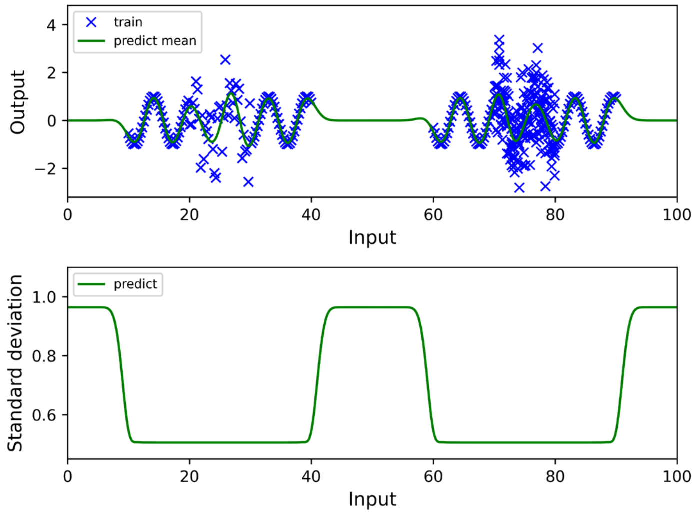
$\mathbb{E}$ -> Expectation
$\mathbb{V}$ -> Variance
R -> Number of outputs

Issues with using multiple reference outputs in GP

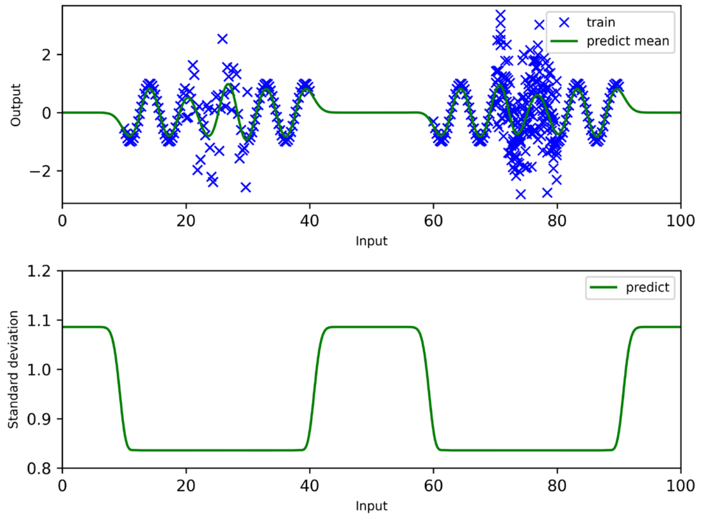
- Standard GP does not have capacity to learn data uncertainty.

$$p(\boldsymbol{y}_1, \cdots, \boldsymbol{y}_R | \boldsymbol{x}) \propto \mathcal{N}\left(\mathbb{E}_{r=1}^{R}(\boldsymbol{y}_r); \boldsymbol{0}, \boldsymbol{K}(\boldsymbol{x}, \boldsymbol{x}) + \frac{\sigma^2}{R}\boldsymbol{I}\right) \mathcal{N}\left(\sqrt{\mathbb{V}_{r=1}^{R}(\boldsymbol{y}_r)}; \boldsymbol{0}, \frac{\sigma^2}{R}\boldsymbol{I}\right)$$

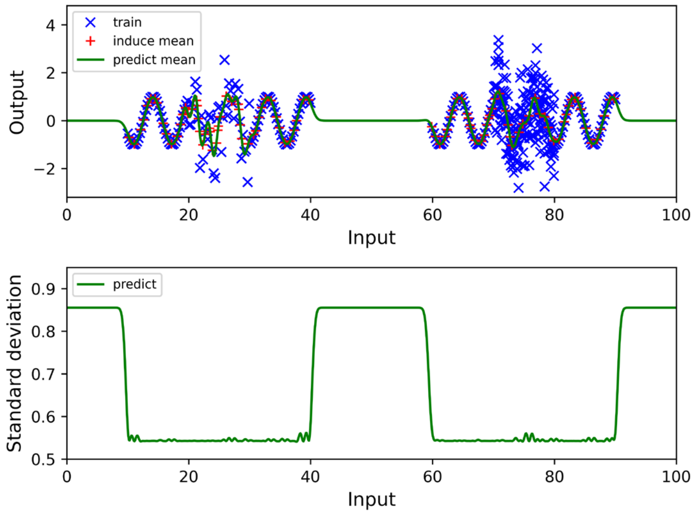
- Standard training criteria do not encourage GP to learn data uncertainty.



Standard GP



GP with multiple training outputs



Variational GP with multiple training outputs

ARES PUBLIC

$\boldsymbol{m}, \boldsymbol{S}$ -> Variational parameters
$\boldsymbol{u}$ -> Inducing latent variables
$\boldsymbol{z}$ -> Inducing inputs

Variational approximation

- Variational approximation:

$$p(\hat{f}|\boldsymbol{y}, \hat{x}, \boldsymbol{x}) \approx \int p(\hat{f}|\hat{x}, \boldsymbol{u}, \boldsymbol{z})\mathcal{N}(\boldsymbol{u}; \boldsymbol{m}, \boldsymbol{S})d\boldsymbol{u}$$

- Approximate posterior:

$$p(\hat{y}|\boldsymbol{y}, \hat{x}, \boldsymbol{x}) \approx \mathcal{N}\left(\hat{y}; \boldsymbol{a}^{\mathrm{T}}\boldsymbol{m}, k(\hat{x}, \hat{x}) + \boldsymbol{a}^{\mathrm{T}}[\boldsymbol{S} - \boldsymbol{K}(\boldsymbol{z}, \boldsymbol{z})]\boldsymbol{a} + \sigma^2\right)$$
$$\boldsymbol{a} = \boldsymbol{K}(\boldsymbol{z}, \boldsymbol{z})\boldsymbol{k}(\boldsymbol{z}, \hat{x})$$

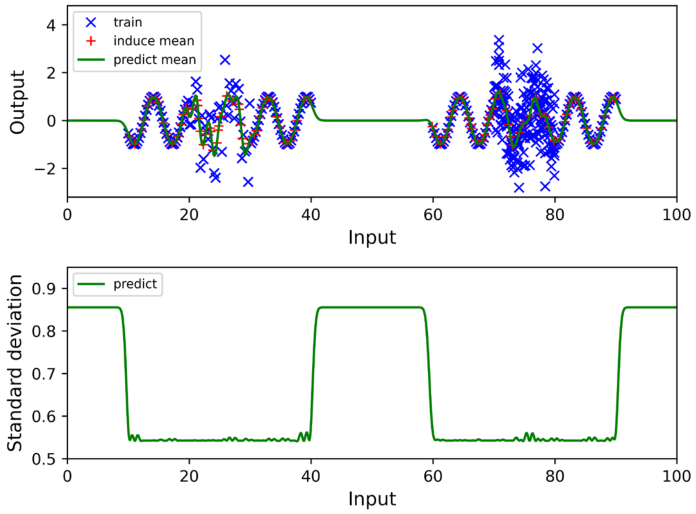
- Able to learn about data uncertainty into $\boldsymbol{a}^{\mathrm{T}}[\boldsymbol{S} - \boldsymbol{K}(\boldsymbol{z}, \boldsymbol{z})]\boldsymbol{a}$.

- Variational approximation originally allows for:
  - Non-Gaussian output density functions.
  - Mini-batch training.

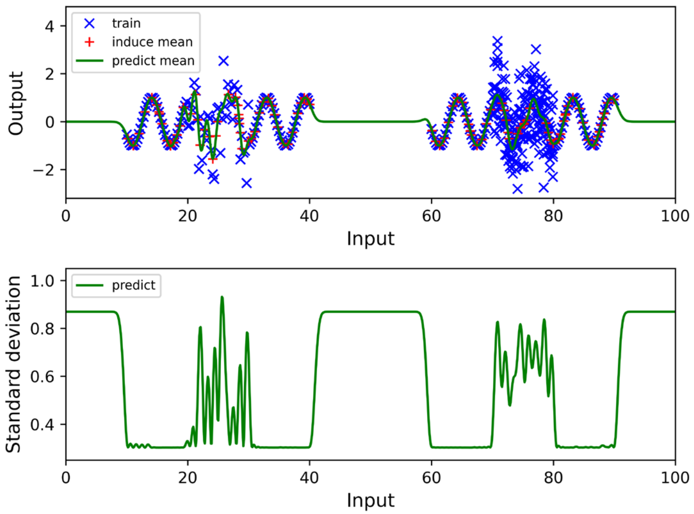
J. Hensman, A. de G. Matthews, and Z. Ghahramani, *"Scalable variational Gaussian process classification,"* AISTATS, 2015

Train GP to capture data uncertainty

- Train GP by minimising distance to reference output distribution.



Variational GP with
multiple training outputs



Minimise distance
between distributions

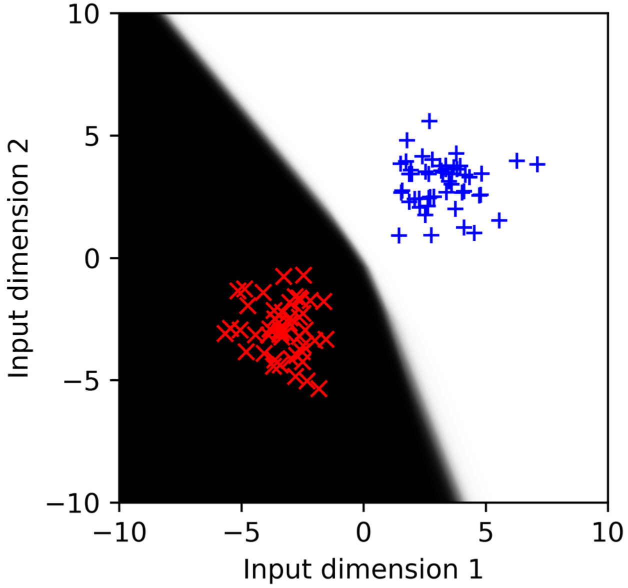
Learning distributional uncertainty from a Gaussian process

ARES PUBLIC

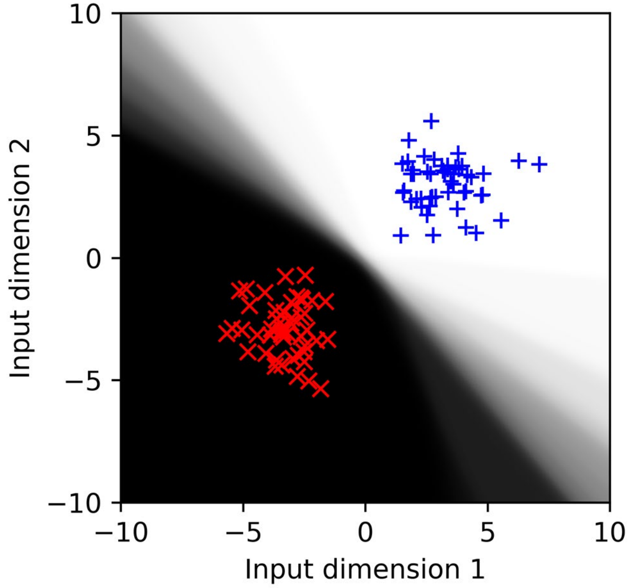
Institute for Infocomm Research

I²R

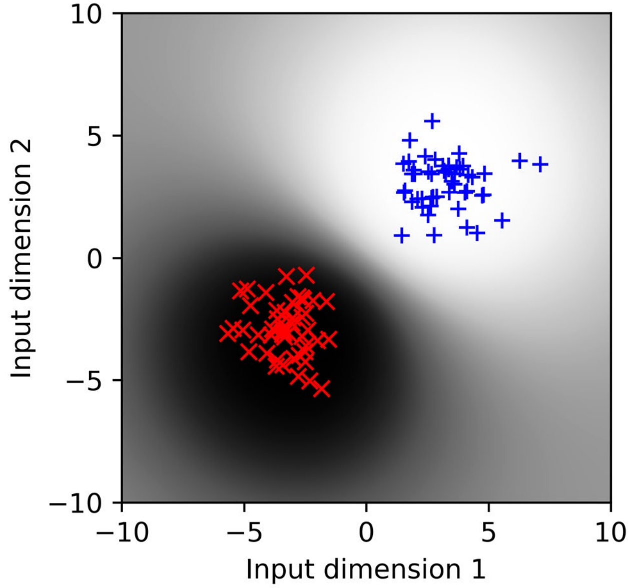
Distributional uncertainty



NN

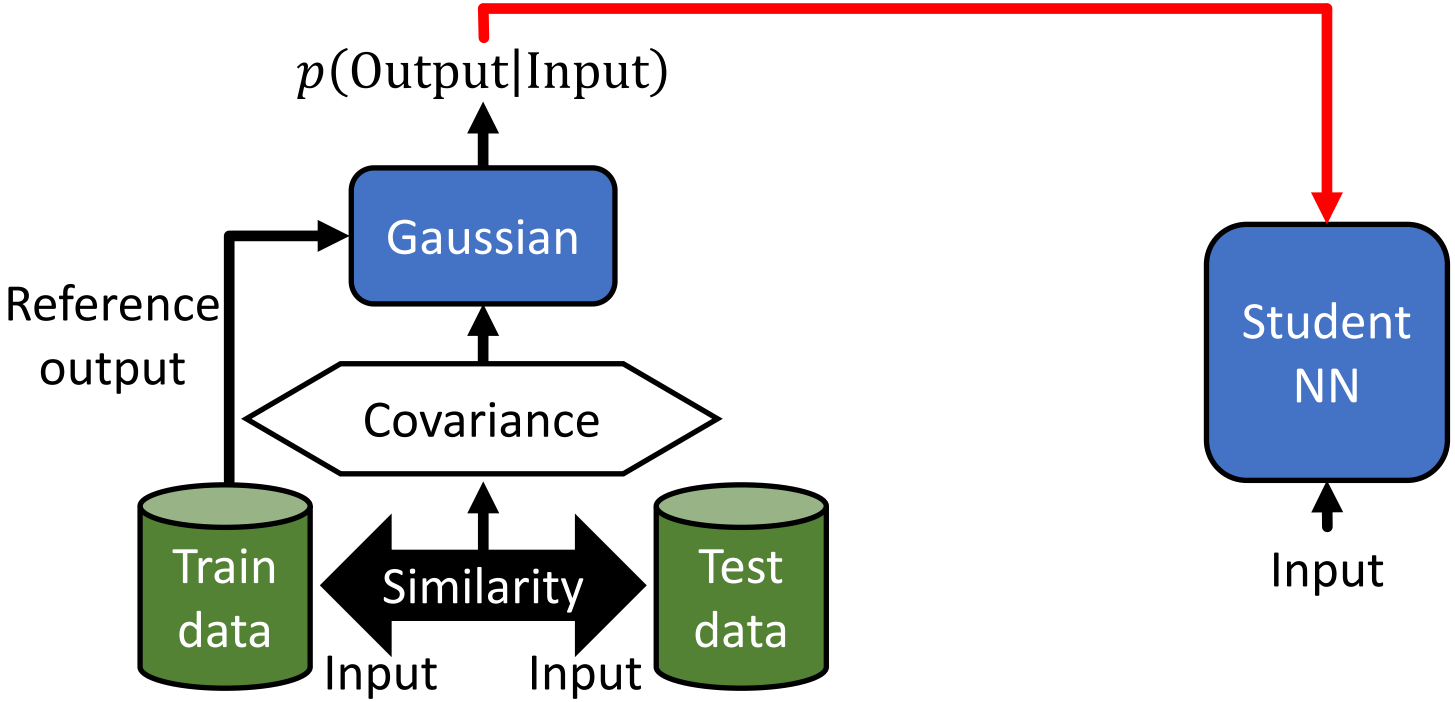


Ensemble of NNs



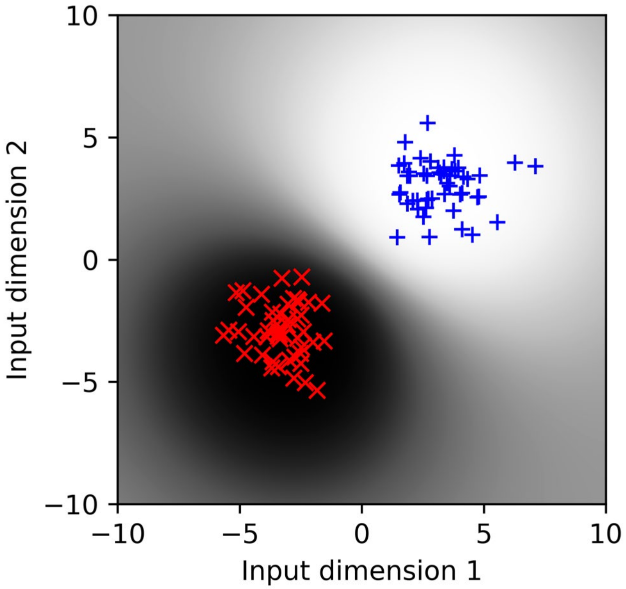
GP

Knowledge distillation

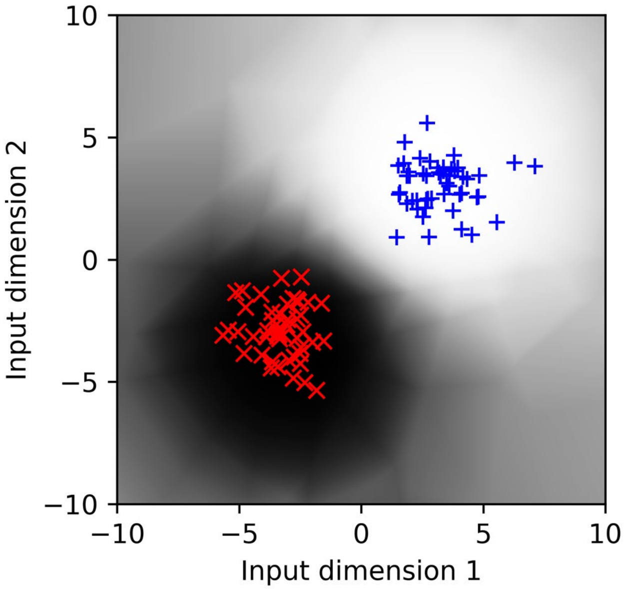


J. Wong, H. Zhang, and N. Chen, *"Distilling knowledge from Gaussian process teacher to neural network student,"* Interspeech, Sep 2023

Knowledge distillation



GP



Student NN

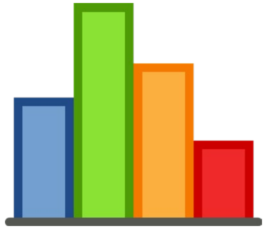
Improving model assumptions

$\mathcal{B}$ -> Beta density function
α, β -> parameters
Γ -> Gamma function

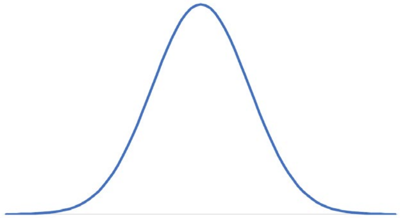
Bounded score range

- Standard model uses softmax, Gaussian, or scalar output.
- Preferred properties:
  - Bounded range of outputs (softmax: yes, Gaussian: no, scalar: maybe)
  - Probabilistic output (softmax: yes, Gaussian: yes, scalar: no)
  - Monotonicity (softmax: no, Gaussian: yes, scalar: yes)
- Using a beta density satisfies all 3 properties.

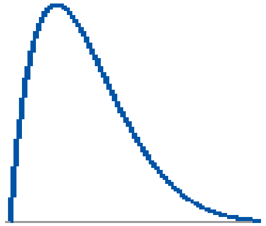
$$\mathcal{B}(y;\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} y^{\alpha-1}(1-y)^{\beta-1}$$



Categorical

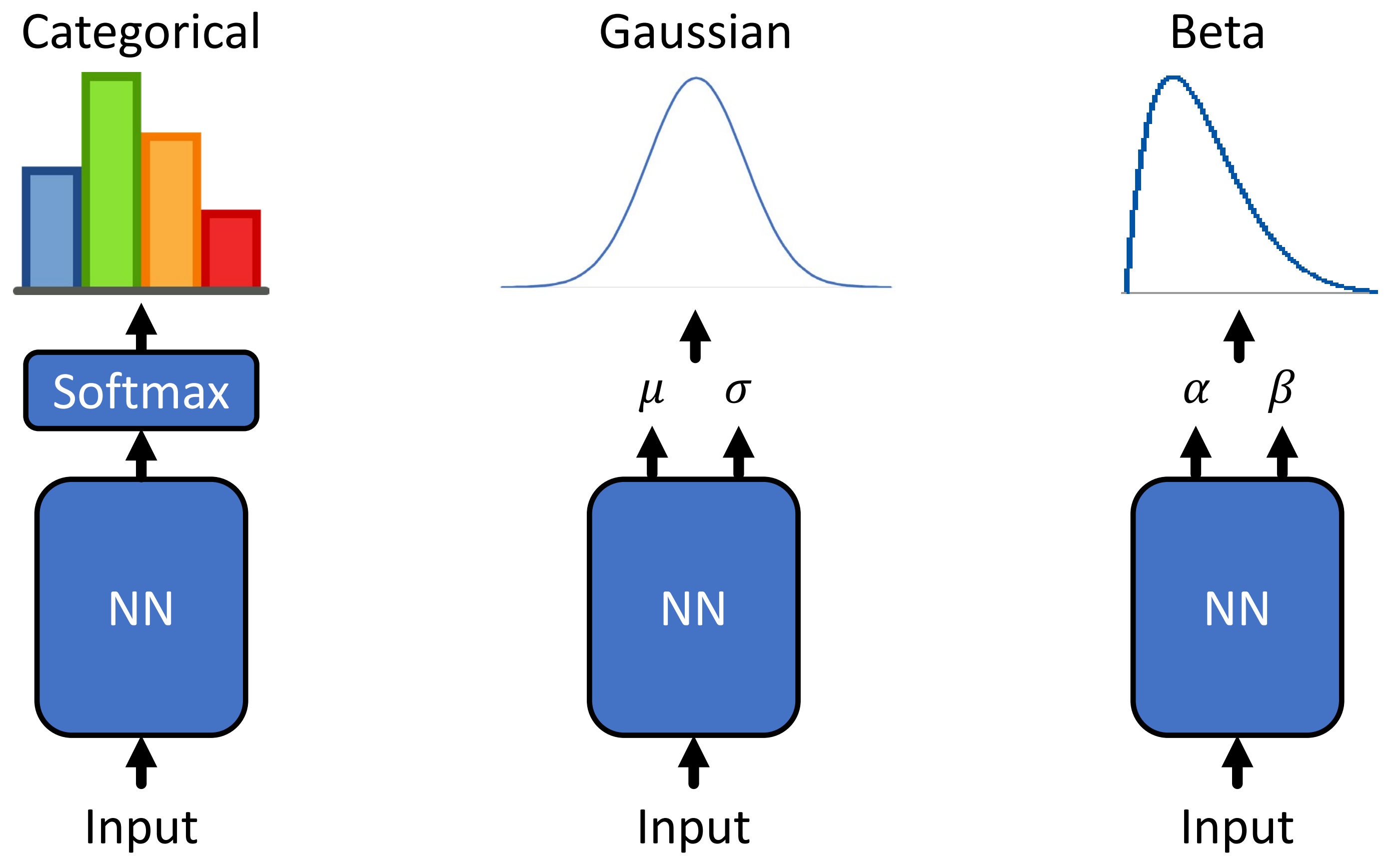


Gaussian



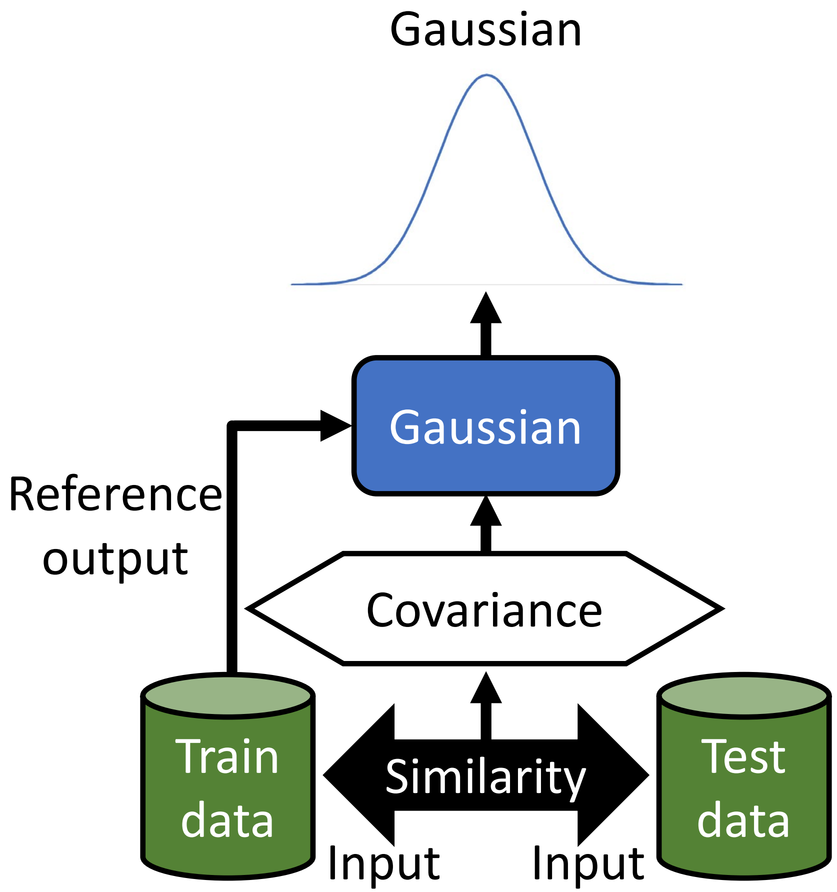
Beta

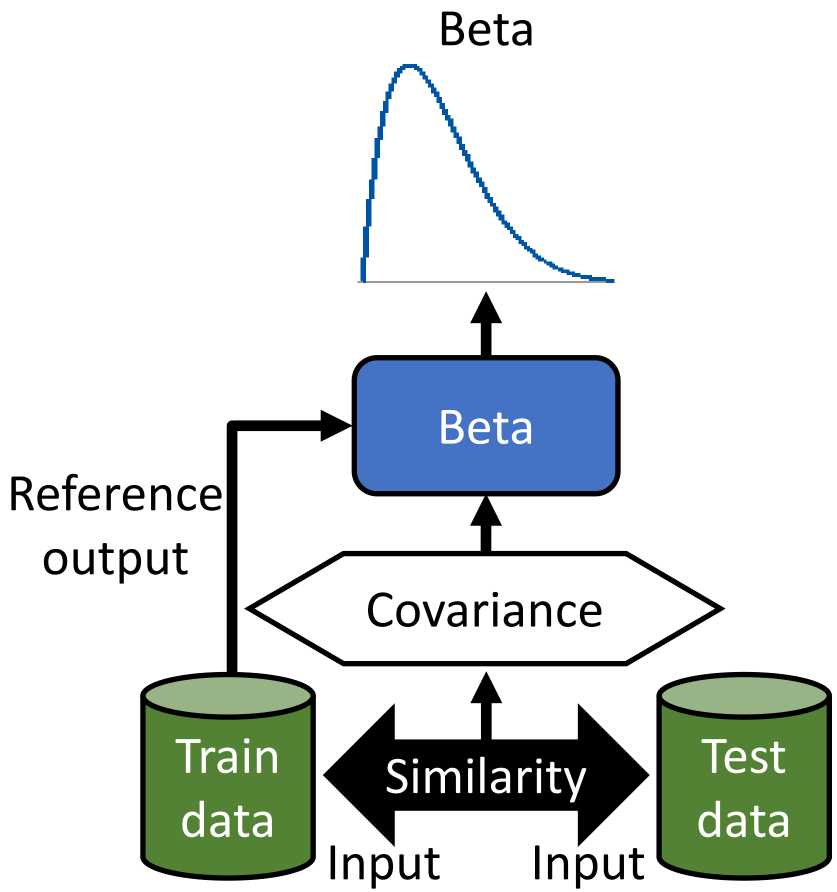
Beta-output neural network



J. Wong, H. Zhang, and N. Chen, *"Modelling inter-rater uncertainty in spoken language assessment,"* IEEE Transactions on Audio, Speech, and Language Processing, vol. 31, Jul 2023

Beta-output Gaussian process





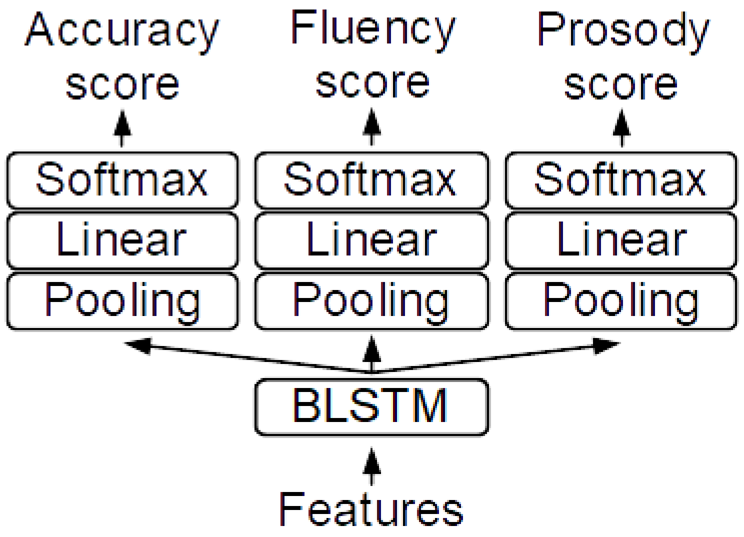
- Implement beta-GP using variational approximation.

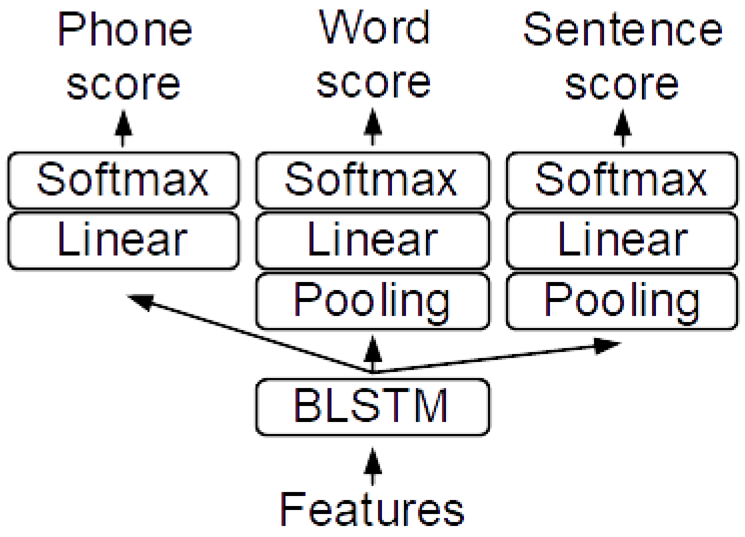
B. Jensen, J. Nielsen, and J. Larsen, *"Bounded Gaussian process regression,"* International Workshop on Machine Learning for Signal Processing, 2013

Other work

Multi-task learning

- Dataset for spoken language assessment is annotated with multiple score types at multiple levels.
- Learn from all score types and levels together.





Y. Gong, Z. Chen, I.-H. Chu, P. Chang, and J. Glass, *"Transformer-based multi-aspect multi-granularity non-native English speaker pronunciation assessment,"* ICASSP, 2022

J. Wong, H. Zhang, and N. Chen, *"Variations of multi-task learning for spoken language assessment,"* Interspeech, 2022

CREATING GROWTH, ENHANCING LIVES

Institute for
Infocomm Research
I²R

MEMORIES FOR THE FUTURE
2002 - 2022

THANK YOU

www.a-star.edu.sg