Deep Learning for Time Series

Session 3b: Practical aspects of forecasting

Romain Tavenard

Multi-step ahead forecasting

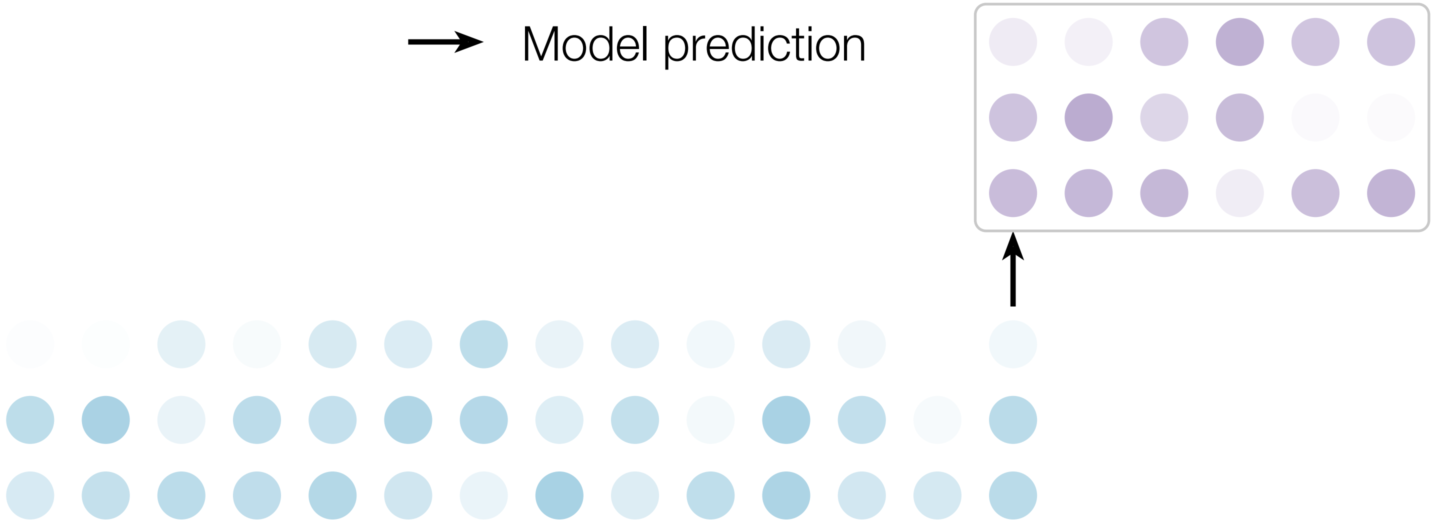
The challenge

- Single-step forecasting: predict x_{t+1} given $x_1, ..., x_t$
- Multi-step forecasting: predict $x_{t+1}, ..., x_{t+H}$ given $x_1, ..., x_t$
- Two main approaches:
  1. Direct forecasting: predict all H steps at once
  2. Autoregressive forecasting: iteratively predict one step at a time

Direct vs Autoregressive

Direct forecasting

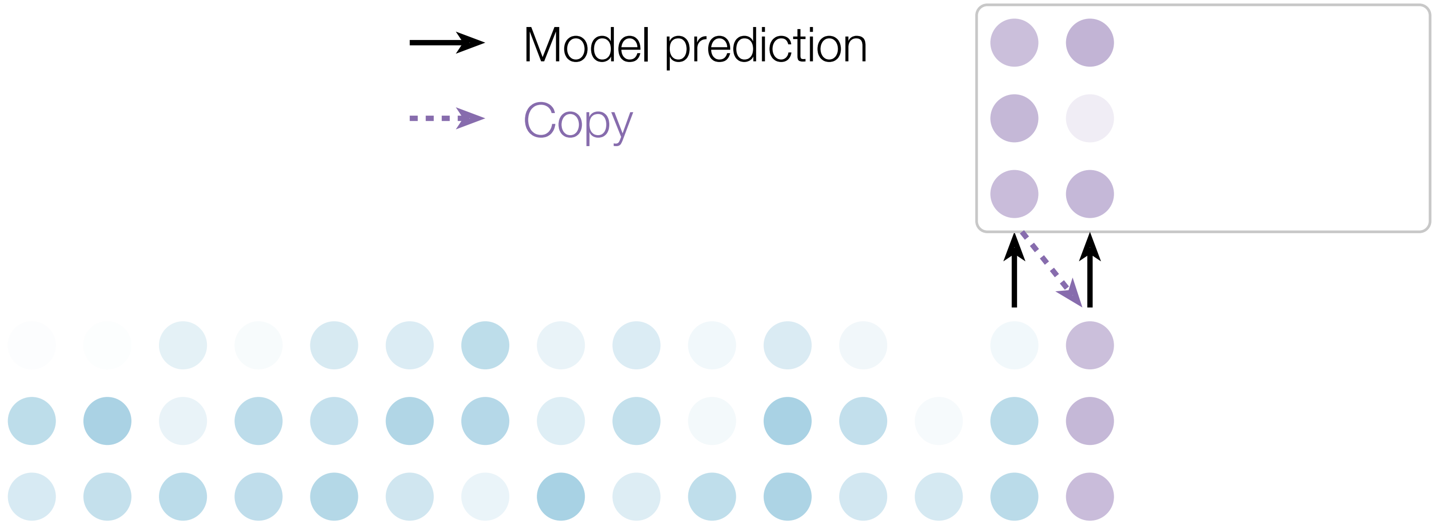
- Model outputs all H future values in one forward pass



- Pros: No error accumulation, faster inference
- Cons: Model must learn to predict all horizons simultaneously

Autoregressive forecasting

- Model predicts $\hat{x}_{t+1}$, then uses it to predict $\hat{x}_{t+2}$, etc.



- Pros: Reuses same model, naturally handles variable horizons
- Cons:
  - Error accumulation, **exposure bias** (train/test mismatch)
  - Model needs to predict all modalities

Training strategies for AR case

Strategy 1: Next-step training / AR inference (*aka* teacher forcing)

- **Training**: Train model to predict one step ahead
  $\mathcal{L} = \frac{1}{T} \sum_t \|x_{t+1} - \hat{x}_{t+1}\|^2$
- **Inference**: Use model autoregressively for multi-step prediction
- Advantage: Simple training procedure
- Disadvantages:
  - Train/test mismatch: model sees ground truth during training but its own predictions during inference
  - Error accumulation: prediction errors compound over time
  - Model may not learn to handle its own prediction errors

Strategy 2: Curriculum learning

- Gradually increase the autoregressive window during training
- Start with short sequences, progressively increase to full horizon
- Typical training procedure:
  1. Epoch 1-10: predict 1 step ahead
  2. Epoch 11-20: predict 2 steps ahead (autoregressively)
  3. … continue until reaching full horizon H
- Benefits:
  - Model learns to handle its own predictions progressively
  - Reduces exposure bias by gradually exposing model to autoregressive inference

Training strategies for AR case

Strategy 3: Scheduled sampling

- During training, randomly replace ground truth with model predictions
- Probability of using prediction increases over time
- Helps bridge train/test gap

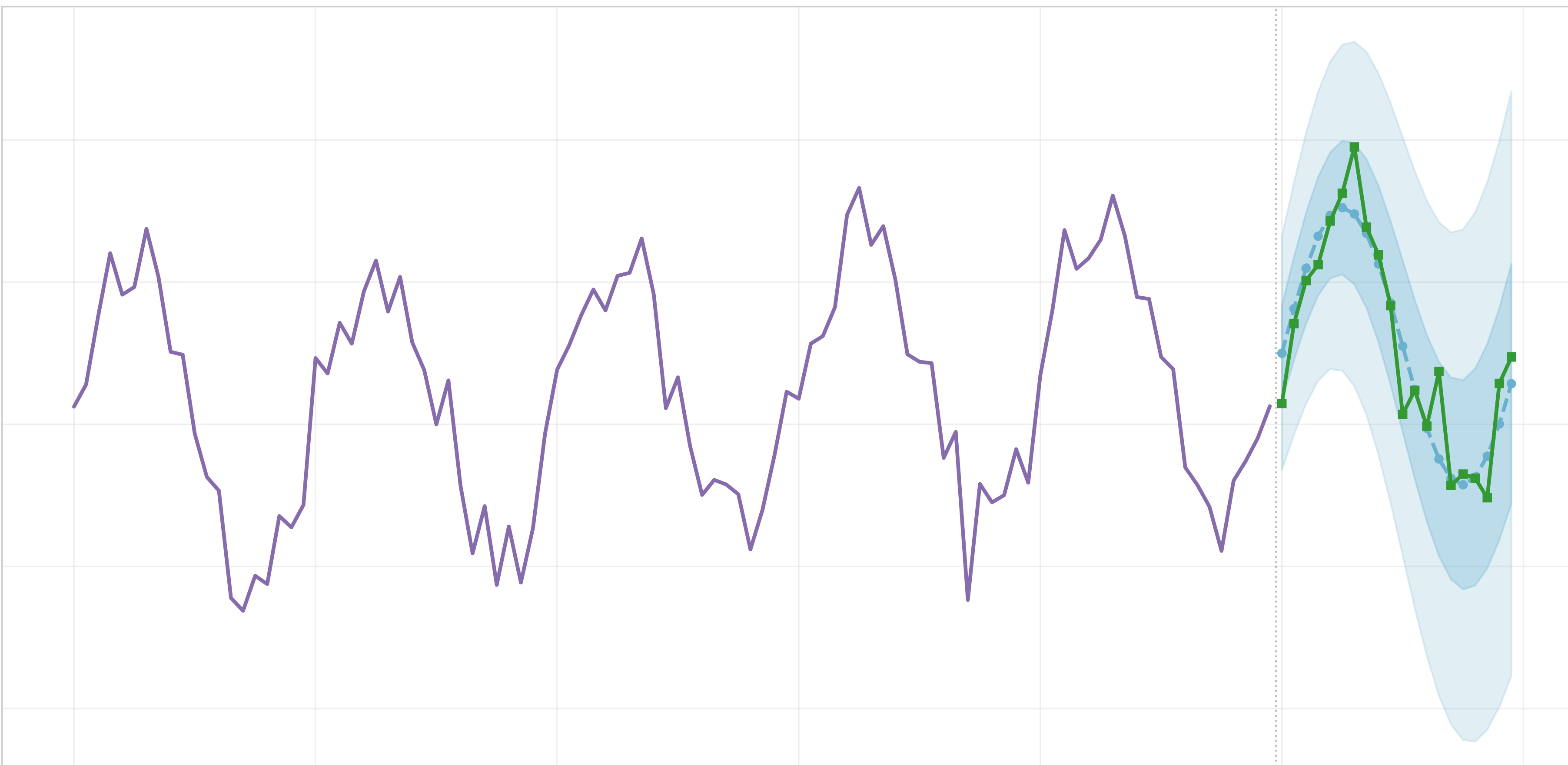
$$x_{t+1}^{\text{input}} = \begin{cases} x_{t+1} & \text{with probability} \quad p \\ \hat{x}_{t+1} & \text{with probability} \quad 1-p \end{cases}$$

where p decreases from 1 to 0 during training

Probabilistic forecasting

Beyond point predictions

- So far: deterministic models output single values $\hat{x}_{t+h}$
- Reality: future is uncertain
- **Probabilistic forecasting**: predict distribution $p(x_{t+h} \mid x_1, \dots, x_t)$



Why probabilistic forecasting?

- Quantify uncertainty: know when model is confident vs uncertain
- Better decision making: risk-aware planning
- Handle non-Gaussian error distributions: capture skewness

Approaches to probabilistic forecasting

1. Parametric distributions

- Model outputs parameters of a distribution (e.g., mean and variance for Gaussian)
  - $\hat{\mu}_{t+h}, \hat{\sigma}_{t+h} = f(x_1, ..., x_t)$
  - Predict $x_{t+h} \sim \mathcal{N}(\hat{\mu}_{t+h}, \hat{\sigma}^2_{t+h})$

2. Quantile regression

- Predict multiple quantiles (e.g., 10th, 50th, 90th percentiles)
- Captures uncertainty without distributional assumptions

Parametric probabilistic models

- Model outputs distribution parameters
- Training via maximum likelihood
- Example: Negative Gaussian log-likelihood

$$
\begin{aligned}
\ell &= -\log p\left(x_{t+h} \mid \hat{\mu}_{t+h}, \hat{\sigma}_{t+h}\right) \\
&= \frac{1}{2} \log\left(2\pi\hat{\sigma}^2_{t+h}\right) + \frac{\left(x_{t+h} - \hat{\mu}_{t+h}\right)^2}{2\hat{\sigma}^2_{t+h}}
\end{aligned}
$$

- Model learns both mean prediction and uncertainty

Quantile regression

- Predict multiple quantiles simultaneously
- Loss function: quantile loss (*aka* pinball loss)

$$\ell_\tau(x_{t+h}, \hat{x}_{t+h}) = \begin{cases} \tau \cdot (x_{t+h} - \hat{x}_{t+h}) & \text{if} \quad \hat{x}_{t+h} \leq x_{t+h} \\ (\tau - 1) \cdot (x_{t+h} - \hat{x}_{t+h}) & \text{if} \quad \hat{x}_{t+h} > x_{t+h} \end{cases}$$

- Common quantiles: $\tau \in \{0.1, 0.5, 0.9\}$
  - Median ($\tau = 0.5$) provides point forecast
  - Other quantiles provide uncertainty intervals

Evaluation of probabilistic forecasts

Calibration

- Predicted 90% intervals should contain true value ~90% of the time

Sharpness

- Among well-calibrated forecasts, prefer narrower intervals

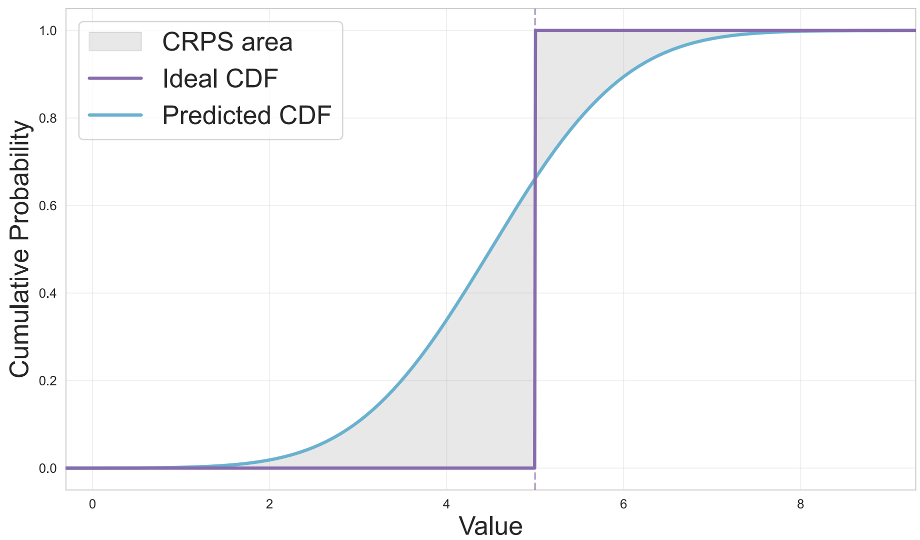
Proper scoring rules

- Log-likelihood on held-out data
- Continuous Ranked Probability Score (CRPS)

Evaluation of probabilistic forecasts

Continuous Ranked Probability Score (CRPS)

- Generalizes MAE to probabilistic forecasts
- Main idea: how close is the CDF of the predicted distribution one of the ideal distribution (a Dirac at the true value)?
    - CRPS: Area between the two CDFs



Summary

- Multi-step forecasting strategies:
  - Direct: predict all steps at once
  - Autoregressive: iterative one-step predictions
- Training considerations:
  - Next-step training simple but suffers from exposure bias
  - Curriculum learning helps bridge train/test gap
- Probabilistic forecasting:
  - Captures uncertainty in predictions
  - Multiple approaches: parametric, quantile regression
  - Important for real-world decision making