Introduction

- DeepAR, a methodology for producing accurate probabilistic forecasts, based on training an auto-regressive recurrent network model on a large number of related time series
- **overcome many of the challenges faced by widely-used classical approaches**
  - forecasting thousands or millions of related time series.
  - alleviate the time and labor intensive manual feature engineering and model selection steps required by classical techniques.
  - provide forecasts for items with little or no history at all
  - can incorporate a wide range of likelihood functions, allowing the user to choose one that is appropriate for the statistical properties of the data.

Model

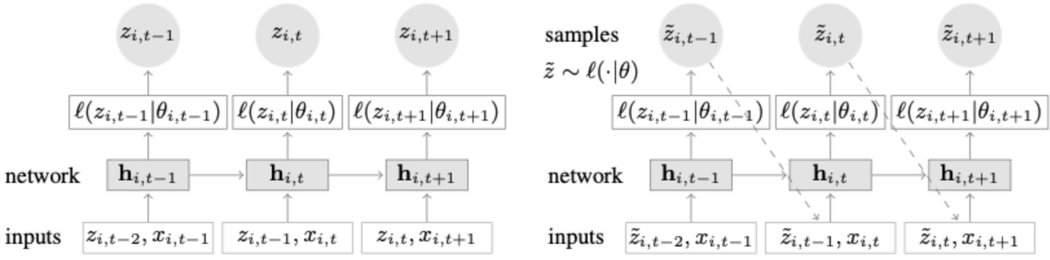
Denoting the value of time series i at time t by $z_{i,t}$, our goal is to model the conditional distribution

$$P\left(\mathbf{z}_{i,t_0:T}|\mathbf{z}_{i,1:t_0-1},\mathbf{x}_{i,1:T}\right)$$

assume

$$Q_\Theta\left(\mathbf{z}_{i,t_0:T}|\mathbf{z}_{i,1:t_0-1},\mathbf{x}_{i,1:T}\right)=\prod_{t=t_0}^{T}Q_\Theta\left(z_{i,t}|\mathbf{z}_{i,1:t-1},\mathbf{x}_{i,1:T}\right)=\prod_{t=t_0}^{T}\ell\left(z_{i,t}|\theta\left(\mathbf{h}_{i,t},\Theta\right)\right)$$

$$\mathbf{h}_{i,t}=h\left(\mathbf{h}_{i,t-1},z_{i,t-1},\mathbf{x}_{i,t},\Theta\right)$$



LikeliHood Model

- The likelihood $l(z|\theta)$ determines the “noise model”, and should be chosen to match the statistical properties of the data.

Gaussian likelihood

$$\ell_{\mathrm{G}}(z|\mu,\sigma)=\left(2\pi\sigma^2\right)^{-\frac{1}{2}}\exp\left(-(z-\mu)^2/\left(2\sigma^2\right)\right)$$
$$\mu\left(\mathbf{h}_{i,t}\right)=\mathbf{w}_\mu^T\mathbf{h}_{i,t}+b_\mu \text{ and } \sigma\left(\mathbf{h}_{i,t}\right)=\log\left(1+\exp\left(\mathbf{w}_\sigma^T\mathbf{h}_{i,t}+b_\sigma\right)\right)$$

Negative binomial

$$\ell_{\mathrm{NB}}(z|\mu,\alpha)=\frac{\Gamma\left(z+\frac{1}{\alpha}\right)}{\Gamma(z+1)\Gamma\left(\frac{1}{\alpha}\right)}\left(\frac{1}{1+\alpha\mu}\right)^{\frac{1}{\alpha}}\left(\frac{\alpha\mu}{1+\alpha\mu}\right)^{z}$$
$$\mu\left(\mathbf{h}_{i,t}\right)=\log\left(1+\exp\left(\mathbf{w}_\mu^T\mathbf{h}_{i,t}+b_\mu\right)\right) \quad \text{and} \quad \alpha\left(\mathbf{h}_{i,t}\right)=\log\left(1+\exp\left(\mathbf{w}_\alpha^T\mathbf{h}_{i,t}+b_\alpha\right)\right)$$

Training

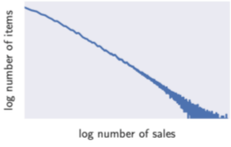
Maxmizing

$$\mathcal{L}=\sum_{i=1}^{N}\sum_{t=t_0}^{T}\log\ell\left(z_{i,t}|\theta\left(\mathbf{h}_{i,t}\right)\right)$$

- 滑窗来确定 training instances
- 保证预测部分都有真实值
- may chose $t=1$ to lie before the start of the time series, e.g. 2012-12-01 in the example above, padding the unobserved target with zeros. This allows the model to learn the behavior of “new” time series taking into account all other available features.

Scale

销售数据：



Scale factor

$$\nu_i=1+\frac{1}{t_0}\sum_{t=1}^{t_0}z_{i,t}$$

- NN has a limited operating range, **input**: $z_{i,t}/\nu_i$ **output**: $\mu=\nu_i\log\left(1+\exp\left(o_\mu\right)\right)$ and $\alpha=\log\left(1+\exp\left(o_\alpha\right)\right)/\sqrt{\nu_i}$
- 由于数据不平衡，均匀随机抽样难以学习到销售量很大的时间序列的特点，因此抽样的概率与 scale v_i 成正比

Features

- item-dependent or time-dependent
- “age” feature, i.e., the distance to the first observation in that time series
- day-of-the-week and hour-of-the-day for hourly data
- week-of-year for weekly data
- month-of-year for monthly data
- category feature embedding