ESE538 Assignment 3

Forecasting in Time Series: A Mathematical Formulation

Exercises

skip for midterm

1. (4 points) Consider the following Moving Average of order 1, MA(1), process:

$$Y_k = \epsilon_k + \theta\epsilon_{k-1}, \quad \text{with } \epsilon_k \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma_\epsilon^2),$$

where μ is a constant, and θ is the moving average coefficient.

Answer the following questions:

(a) Derive the one-step-ahead forecast density $f_{Y_{k+1}|\mathcal{F}_k}(y_{k+1} \mid \mathcal{F}_k)$, where $\mathcal{F}_k = \{Y_k = y_k, Y_{k-1} = y_{k-1}, Y_{k-2} = y_{k-2}, \dots\}$. *Hint*: The information set does not contain the values of the noise terms.

(b) Derive the value of the optimal forecast $\widehat{y}^\star_{k+1}$ when the loss function is the squared error.

(c) Derive the value of the optimal forecast $\widehat{y}^\star_{k+1}$ when the loss function is the absolute error.

(d) Derive the **two-steps-ahead forecast density**, $f_{Y_{k+2}|\mathcal{F}_k}(y_{k+2} \mid \mathcal{F}_k)$.

(e) Derive the value of the two-steps-ahead optimal forecast $\widehat{y}^\star_{k+2}$ when the loss function is the squared error.

(a):

$Y_{k+1} = \epsilon_{k+1} + \theta\epsilon_k$

Given $\mathcal{F}_k$, we know Y_k, but ϵ_{k+1} is future shock,

So then we have no direct info about ϵ_{k+1},

$\Rightarrow$ it will be expected, which is i.i.d.

$\Rightarrow$ $Y_{k+1} | \mathcal{F}_k \sim \mathcal{N}(\mu + \theta\epsilon_k, \sigma_\epsilon^2)$

the parametric form of one-step-ahead forecast density is: $f_\theta(\hat{y}_{k+1}; y_{1:k}) = \frac{1}{\sqrt{2\pi\sigma_\epsilon^2}} \exp\left(-\frac{(y_{k+1} - (\mu + \theta(y_k - \mu)))^2}{2\sigma_\epsilon^2}\right)$

(b): $SEL = (Y_{k+1} - \hat{y}_{k+1})^2$

$\therefore \hat{y}^*_{k+1} = E[Y_{k+1} | \mathcal{F}_k] = \mu + \theta(y_k - \mu)$ from (a).

(c) $AEL = |Y_{k+1} - \hat{y}_{k+1}|$, which is median of Y_{k+1}

As Y_{k+1} is normally distributed, mean & median are same

$\Rightarrow$ $\hat{y}^*_{k+1} = \mu + \theta(y_k - \mu)$ are unchanged

(d):

2. (6 points) Consider the following Autoregressive Moving Average process, denoted by ARMA(1,1) and defined as:

$$Y_k = \phi Y_{k-1} + \epsilon_k + \theta \epsilon_{k-1}, \quad \text{with } \epsilon_k \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_\epsilon^2) \text{ and } Y_0 = 0,$$

where ϕ and θ are the autoregressive and moving average coefficients, respectively. Assume that, at time k, you have access to the following information set: $\mathcal{F}_k = \{Y_k = y_k, Y_{k-1} = y_{k-1}, Y_{k-2} = y_{k-2}, \dots\}$. Note that the information set does **not** contain the values of the noise terms ϵ_k.

Answer the following questions:

skip for midterm

(a) Derive the one-step-ahead forecast density $f_{Y_{k+1}|\mathcal{F}_k}(y_{k+1} \mid \mathcal{F}_k)$ for the ARMA(1,1) process.

(b) Write down an explicit integral expression for the conditional risk for the squared error (SE) loss function. Your expression should be an integral where the integrand should be an explicit function of the parameters ϕ, θ, σ_ϵ^2, the prediction $\widehat{y}_{k+1}$, the values in the information set $\mathcal{F}_k$, and the integration variable y. *Hint*: The constant $\pi \approx 3.14$ should appear in your integral.

(c) Use the properties of the Gaussian density to derive an explicit expression for the conditional risk for the SE loss function that does not involve integrals. The expression should be a function of the parameters ϕ, θ, σ_ϵ^2, the prediction $\widehat{y}_{k+1}$, and the values in the information set $\mathcal{F}_k$.

(a):

(d) Minimize the conditional risk function with respect to $\widehat{y}_{k+1}$ as follows: Find the derivative of the conditional risk with respect $\widehat{y}_{k+1}$, force the derivative to be zero, and find the value $\widehat{y}_{k+1}^*$ that makes the derivative equal to zero.

(e) What is the relationship between this value and $\mathbb{E}[Y_{k+1} \mid \mathcal{F}_k]$? Derive an expression for this conditional expectation to verify your claim.

3. (4 points) Consider the following AR(1) model:

$$Y_k = \theta Y_{k-1} + \epsilon_k, \quad \text{with } \epsilon_k \overset{\text{iid}}{\sim} \mathcal{N}(0,1) \text{ and } Y_0 = y_0.$$

where θ is the unknown parameter we want to estimate, and the noise variance is fixed at $\sigma_\epsilon^2 = 1$. You are provided with L observations of the time series $y_{1:L} = (y_1, y_2, \ldots, y_L)$.

Answer the following questions:

(a) Write down the likelihood function $\mathscr{L}(\theta; y_{0:L})$ for the AR(1) model using the causal factorization in (3).

(b) Derive the log-likelihood function, defined as $\log \mathscr{L}(\theta; y_1, y_2, \ldots, y_L)$, where $\log(\cdot)$ is the natural logarithm.

(c) Compute the derivative of the log-likelihood function with respect to θ.

(d) Set the derivative equal to zero and solve for θ to obtain the **Maximum Likelihood Estimate (MLE)** $\theta^\star$ as a function of the observations.

(e) Interpret the MLE expression in terms of the autocorrelation of the observed data.

(a): $\mathscr{L}(\theta; y_{0:L}) = P(Y_1 = y_1, Y_2 = y_2, \ldots Y_L = y_L \mid Y_0 = y_0, \theta)$

while for AR(1), $f_\theta(y_{k+1}; y_{1:k}) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(y_{k+1} - \theta y_k)^2\right)$

So by (3), $\mathscr{L}(\theta; y_{0:L}) = \prod_{k=1}^{L} \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}(y_k - \theta y_{k-1})^2\right]$

(b):

$$\log(\mathscr{L}(\theta; y_{0:L})) = L \log\left(\frac{1}{\sqrt{2\pi}}\right) + \sum_{k=1}^{L} \left(-\frac{(y_k - \theta y_{k-1})^2}{2}\right)$$

$$= -\frac{L}{2}\log 2\pi - \frac{1}{2}\sum_{k=1}^{L}(y_k - \theta y_{k-1})^2$$

(c):

$$\frac{\partial \log(\mathscr{L}(\theta; y_{0:L}))}{\partial \theta} = 0 - \frac{1}{2}\sum_{k=1}^{L} \frac{d}{d\theta}(y_k - \theta y_{k-1})^2$$

$$= -\frac{1}{2}\sum_{k=1}^{L} 2(y_k - \theta y_{k-1})(-y_{k-1})$$

$$= \sum_{k=1}^{L} y_{k-1}(y_k - \theta y_{k-1})$$

(d): let $\frac{\partial \log(\mathscr{L}(\theta; y_{0:L}))}{\partial \theta} = 0$

$$\Rightarrow \sum_{k=1}^{L} y_{k-1}(y_k - \theta y_{k-1}) = 0$$

$$\Rightarrow \sum_{k=1}^{L} y_{k-1} y_k - \theta \sum_{k=1}^{L} y_{k-1}^2 = 0$$

So $\theta^* = \dfrac{\sum_{k=1}^{L} y_{k-1} y_k}{\sum_{k=1}^{L} y_{k-1}^2}$

(e): θ^* is sample autocorrelation at lag 1 when mean is 0

$$\Rightarrow \theta^* \approx \frac{\mathrm{Cov}(y_{k-1}, y_k)}{\mathrm{Var}(y_{k-1})}$$

it provides a numerical estimation based on the observation

4. (5 points) Consider the AR(1) model:

$$Y_k = \theta Y_{k-1} + \epsilon_k, \quad \text{with } \epsilon_k \overset{\text{iid}}{\sim} \mathcal{N}(0,1) \text{ and } Y_0 = y_0.$$

where θ is the unknown parameter we want to estimate, and the noise variance is fixed at $\sigma_\epsilon^2 = 1$. You are provided with L observations of the time series $y_{1:L} = (y_1, y_2, \ldots, y_L)$. Let us consider a prior normal prior distribution for the unknown parameter, i.e., $\theta \sim \mathcal{N}(\mu_0, \sigma_0^2)$, where μ_0 and σ_0 are given parameters.

Answer the following questions:

(a) Assume that at time $k = 1$, you observe a single data point $Y_1 = y_1$ from the AR(1) process. Derive both the likelihood function $f_{Y_1|\theta}$ and the posterior distribution $f_{\theta|Y_{0:1}}$ given this single observation y_1 (and the initial condition $Y_0 = y_0$. *Hint*: Show that the posterior distribution remains Gaussian and derive the updated posterior mean μ_1 and posterior

skip for midterm

5. 5. Answer the following short questions (justify your answers):

(a) What is the most appropriate loss function when the one-step-ahead forecast distribution is heavy-tailed?

(b) What is the most appropriate loss function when the one-step-ahead forecast distribution is Gaussian?

(c) What are the key differences between the model-based approach and the Bayesian model-based approach?

(d) What are the advantages and disadvantages of the traditional machine learning approach compared to the model-based approach?

(e) What are the advantages and disadvantages of the deep learning approach compared to the traditional machine learning approach?

(a): Absolute Error (AE) would be better, as
it ① less sensitive to extreme OOD value than SE
② It's more common to have large deviation outliers in heavy-tailed data

(b) Squared Error (SE), as it:
① SE searchs mean of data distribution
② This forcast optimally under normal distrib
③ SE loss penalizes deviations symmetrically
⇒ fits well with gaussian distribu

(c): ① Parameters in Model-based are fixed
in Bayesian ↳ are random variable with prior
② Bayesian provide a principled mechanism for updating beliefs about the data as new information becomes available,

(d): Traditional ML vs. Model-based

Pros:
- Flexibility: ML can learn complex non-linear features, which is hard to model
- Feature Engineering: we can use diverse features to enhance performance

Cons:
- Data & Time Cost: ML requires extra datas to train which may be resource consuming
- Interpretability: ML are usually black-box model, while model-based decision is clearer

5° DL vs ML

Pros

- Much stronger performance: with more complex NN, the forcast accuracy raises quite a lot to solve comple patten
- Auto Feature Extraction: data-driven method can extract spatiotemporal data easily

Cons:

① Overfit: for small data, DL are prone to overfit

② Computation: needs more computing resources than ML

(f) What is the main limitation of Recurrent Neural Networks (RNNs)?

(g) In what scenarios are autoencoders particularly useful?

(h) How can long-term dependencies be incorporated into Convolutional Neural Networks (CNNs)?

(i) What are the advantages of Transformers compared to RNNs?

(f): Vanishing gradient problem. RNNs can train to nothing, the gradient tend to explode/vanish during back propagation and RNNs can not retain info over long-term sequence

(g):

⇒ Dimensionality reduction or Representation Learning

reconstruction ← VAE ← Decoding

⇒ very good for anomaly detection

(h): Dilated convolutions

⇒ filter inputs spaced by gaps,

⇒ pass wider data without increasing depth & #Param

(i): ① Parallel Processing: faster train & inference

② global self-attention mechanism enables long-term context learning

⇒ avoid Vanishing gradient problem.

6. (From 1A—3 points) Consider a Markov chain $\mathcal{X} = \{X_1, X_2, \ldots\}$ with state space $\mathcal{S} = (1, 2, 3)$ and transition matrix:

$$P = \begin{bmatrix} 0 & 0.8 & * \\ 0.1 & * & 0.3 \\ 0.2 & 0.8 & * \end{bmatrix}.$$

(a) Fill in the unknown entries, denoted by $*$, in the transition matrix.

(b) Find the probability that $X_2 = s_3$ given that $X_0 = s_1$.

(c) Given that $X_0 = s_1$, compute the distribution vector $\boldsymbol{\pi}_2$, representing the state probabilities at time step 2.

(a): By Markov Chain, we have each row sum to 1

$$\Rightarrow \begin{cases} 0 + 0.8 + *_1 = 1 \\ 0.1 + *_2 + 0.3 = 1 \\ 0.2 + 0.8 + *_3 = 1 \end{cases} \Rightarrow \begin{cases} *_1 = 0.2 \\ *_2 = 0.6 \\ *_3 = 0 \end{cases}$$

$$\therefore P = \begin{bmatrix} 0 & 0.8 & 0.2 \\ 0.1 & 0.6 & 0.3 \\ 0.2 & 0.8 & 0 \end{bmatrix}$$

(b): $P(X_2 = S_3 \mid X_0 = S_1)$

$$= \sum_{k=1}^{3} P(X_2 = S_3 \mid X_1 = S_k) \cdot P(X_1 = S_k \mid X_0 = S_1)$$

$$\Rightarrow P(X_2 = S_3 \mid X_0 = S_1)$$

$$= P_{13}^2 = (P_{11} \times P_{13}) + (P_{12} \times P_{23}) + (P_{13} \times P_{33})$$

$$= 0 \times 0.2 + 0.8 \times 0.3 + 0.2 \times 0$$

$$= 0.24$$

(c): $\pi_2 = \pi_0 P^2$

while $\pi_0 = [1, 0, 0]$

So $\pi_2 = [0.12, 0.64, 0.24]^T$

7. (From 1A—3 points) What is the maximum number of non-zero entries in the transition matrix P of a higher-order Markov chain with states $\mathcal{S} = \{s_1, s_2, \ldots, s_H\}$ and memory m? Your answer should be a function of H and m.

An m-th order Markov chain is a stochastic process where the probability of transiting to the next state depends on the previous m states

each state history = m with current state, chain length is m+1

$\Rightarrow H^{m+1}$ possibility

And including the next possible transition to be $\times H$ $\Rightarrow$ Total transition $= H^{m+1} \times H = H^{m+2}$

Thus maximum # non-zero entry of $P = H^{m+2}$

Verify: $H = \{S_1, S_2\}$, $m = 1$

$\Rightarrow (S_1, S_1)\ (S_1, S_2)\ (S_2, S_1)\ (S_1, S_1)$

$\Rightarrow H = 2$ possible next state so 8 in total

$\Rightarrow * = 2^3$

[i. this further verify

8. (From 1A—4 points) Below is a matrix representing the autocovariance matrix of a random process with two components: $\mathcal{Y}^1$ and $\mathcal{Y}^2$. Information about the autocorrelation functions and cross-correlation functions between the processes are shown for lag values less than 3 in the following table:

Autocorrelation: $R_{\mathcal{Y}^1}(h)$		Cross-correlation: $R_{\mathcal{Y}^1\mathcal{Y}^2}(h)$	
Lag	Value	Lag	Value
0	1.0	0	0.8
1	0.5	1	0.4
2	0.2	2	0.2
3	0.1	3	0.1
Cross-correlation: $R_{\mathcal{Y}^2\mathcal{Y}^1}(h)$		Autocorrelation: $R_{\mathcal{Y}^2}(h)$	
Lag	Value	Lag	Value
0	0.8	0	1.0
1	0.3	1	0.4
2	0.2	2	0.1
3	0.05	3	0.05

Assuming that the variance of $\mathcal{Y}^1$ and $\mathcal{Y}^2$ are 1 and 2, respectively, answer the following questions:

(a) Calculate the covariance between $\mathcal{Y}^1$ and $\mathcal{Y}^2$ at lag 0.

(b) Compute $R_{\mathcal{Y}^1}(-2)$

(c) Calculate $R_{\mathcal{Y}^1\mathcal{Y}^2}(-1)$?

(d) How much of the total variance of $\mathcal{Y}^1$ is explained by lagged values up to lag 3?

1a): $Cov(Y^1, Y^2) = R_{y_1y_2}(0) - \mu_{y_1}\mu_{y_2}$

$= 0.8 \times \sqrt{1}\sqrt{2}$

$= 0.8\sqrt{2} \simeq 1.131$

1b): $R_{y^1}(-2) = R_{y^1}(2) = 0.2$

because autocorrelation is symmetric for stationary process

1c): We assume mean = 0 here,

$R_{y^1y^2}(h) = E[Y_k^1 Y_{k-h}^2] = E[Y_{k+h}^1 Y_k^2]$

$= E[Y_k^2 Y_{k-1+h}^1] = R_{y_2y_1}(-h)$

$\therefore$

$R_{y^1y^2}(-1) = R_{y_2y_1}(1) = 0.3$

1d): Since here we looking for amount to Sum of y^1 var to 3

$\Rightarrow$ it's squared sum of autocorrelation at lag 1, 2, 3

$Ans = R_{y^1}^2(1) + R_{y^1}(2)^2 + R_{y^1}(3)^2$

$= 0.5^2 + 0.2^2 + 0.1^2$

$\simeq 30\%$

30% of total variance of y^1