Machine Learning WS2014/15 (Unsupervised Learning)

Lecturer: Prof. Dr. Bethge

2. Directed/Markov chain/autoregressive models

Task 1:
Derive

a) the posterior distribution for a Gaussian prior $p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\mathbf{m}_1, C_1)$ and Gaussian likelihood $p(\mathbf{m}_2|\mathbf{x}) = \mathcal{N}(\mathbf{x}|\mathbf{m}_2, C_2)$

b) the m-dimensional marginal distribution

c) the m-dimensional conditional distribution

of a multivariate n-dimensional Gaussian (see lecture notes).

Remark: You do not have to derive the normalization constant.

Task 2:
Compute the autocorrelation function of the stationary process $x_{t+1} = ax_t + \eta_t$ with $0 < a < 1$, initial condition $x_0 = 0$ and uncorrelated innovations

$$E[\eta_t] = 0, \quad E[\eta_t \eta_{t'}] = \begin{cases} 1 & , t = t' \\ 0 & , t \neq t' \end{cases} .$$

Task 3:
The goal of this task is to fit different directed models, to generate synthetic samples from the different models, and to quantitatively compare their performance.
Download the file `training_data.mat` from Ilias which contains a long time series $(x_1, \ldots, x_T)$ and fit the following types of stationary Markov chain models to the data:

a) m-th order regressive Gaussian process model

$$\hat{p}_{GP}(x_k|x_{k-m}, \ldots x_{k-1}) = \hat{p}_{GP}(x_k|\mathbf{x}_{k,m}) = \mathcal{N}(x_k|\mathbf{w}^\top \mathbf{x}_{k,m} + b, \sigma^2).$$

b) The conditional mean estimate $\hat{E}[x_k] = \mathbf{w}^\top \mathbf{x}_{k,m} + b$ can be interpreted as a prediction (a point estimate) and the difference between the actual sample x_k and $\hat{E}[x_k]$ can be interpreted as prediction error $\epsilon_k := x_k - \hat{E}[x_k]$. The Gaussian process model assumes that $\hat{p}_{GP}(\epsilon_k|\mathbf{x}_{k,m}) = \mathcal{N}(x_k|0, \sigma^2)$ is a constant Gaussian distribution. The simplest generalization from this model is to replace the Gaussian error distribution by some arbitrary distribution that is independent of $\mathbf{x}_{k,m}$. This model could be called m-th order linear least squares conditional mean model with constant non Gaussian error distribution.

$$\hat{p}_{lls}(x_k|x_{k-m}, \ldots x_{k-1}) = \hat{p}_{lls}(x_k|\mathbf{x}_{k,m}) = \hat{p}(\underbrace{x_k - (\mathbf{w}^\top \mathbf{x}_{k,m} + b)}_{=\epsilon_k}).$$

where $\hat{p}$ can be an arbitrary constant probability density function.

c) Another obvious extension of the linear least squares conditional mean model is to assume "heteroskedasticity", that is to make the error distribution dependent on the past $\hat{p}(\epsilon) = \hat{p}(\epsilon|\mathbf{x}_{k,m})$. While it is impossible for a Gaussian process to have a history-dependent variance, a common model known as ARCH process assumes a conditional Gaussian distribution with a linear history-dependence of the conditional variance:

$$\hat{p}_{ARCH}(x_k|x_{k-m},\dots x_{k-1}) = \hat{p}_{ARCH}(x_k|\mathbf{x}_{k,m}) = \mathcal{N}(x_k|\mathbf{w}^\top\mathbf{x}_{k,m}+b, \mathbf{v}^\top\mathbf{x}_{k,m}+c)$$

with $u_j \geq 0,\ \forall j = 1,\dots,m$ and $c \geq 0$.

d) However, it is not necessary to restrict the error distribution to be Gaussian and also the dependence of the conditional variance on the past does not need to be linear. For example, we can extend the m-th order linear least squares conditional mean model in b) with a generalized linear conditional variance dependency

$$\hat{p}_{gen}(x_k|x_{k-m},\dots x_{k-1}) = \hat{p}_{gen}(x_k|\mathbf{x}_{k,m}) = \hat{p}\left(\frac{x_k - (\mathbf{w}^\top\mathbf{x}_{k,m}+b)}{f(\mathbf{v}^\top\mathbf{x}_{k,m}+c)}\right)$$

where $f : R^m \to R$ can be an arbitrary function and $\hat{p}$ can be an arbitrary constant probability density function. A simple way to find a suitable f is by visual inspection when plotting the absolute value of the prediction error $|\epsilon_k|$ against $\mathbf{v}^\top\mathbf{x}_{k,m}+c$.

Try to fit these four different models for $m = 0, 1, 2, 4, 8, 16$ and then download the file `test_data.mat` from Ilias which contains another time series $(y_1,\dots,y_T)$ that is statistically similar to $(x_1,\dots,x_T)$.

- Generate five synthetic samples from the Gaussian process model (a) that you fitted for each $m = 0, 1, 2, 4, 8, 16$ by using the first m time steps as initial condition. Make one plot for each m (= six plots in total, you can use subplot to put them in the same figure) showing the first hundred time steps of the original time-series $(y_1,\dots,y_{100})$ in the test set and superimposed the five samples you sampled from the Gaussian process model.

- Generate five synthetic samples from the other three model (b,c,d) for $m = 1$ and $m = 16$. Make one plot for each model and for the two different m (= six plots in total) again showing the first hundred time steps of the original time-series $(y_1,\dots,y_{100})$ in the test set and superimposed the five samples you sampled from the Gaussian process model.

- Evaluate the log-likelihood of the four different models on both the training set $(x_1,\dots,x_T)$ and on the test set $(y_1,\dots,y_T)$ for all six different values of $m = 0, 1, 2, 4, 8, 16$. Generate two figures (one for the training set and one for the test set) showing the log-likelihood as a function of m for the four different models.

- Use your results on the test set to compute an estimate for the mutual information between present and past samples (take your best 0-th order model as estimate for the marginal entropy). Generate a figure showing the mutual information estimates as a function of m for the four different models.