10-702 Statistical Machine Learning: Assignment 3

Due Friday, February 19

Hand in to Sharon Cavlovich, GHC (Gates Hillman Center) 8215 by 3:00. Use R for all numerical computations.

1. Undirected graphical models

   (a) Let $X = (X_1, \ldots, X_d)^T$ where each $X_j \in \{0, 1\}$. Consider the log-linear model

   $$\log p(x) = \beta_0 + \sum_{j=1}^{d} \beta_j X_j + \sum_{j<k}^{d} \beta_{jk} X_j X_k + \cdots + \sum_{j<k<\ell}^{d} \beta_{jk\ell} X_j X_k X_\ell + \cdots +$$

   Suppose that $\beta_A = 0$ whenever $\{1, 2\} \subset A$. Show that

   $$X_1 \amalg X_2 \,\Big|\, X_3, \ldots, X_d.$$

   Remark: Log-linear models are covered in Section 17.6, however, this question does not require knowledge of any technical details related to the log-linear models.

   (b) Let $X = (X_1, \ldots, X_d)^T$ be a discrete random vector, where each X_i takes values in a finite set $\mathcal{X}$, and let $p(x)$ denote the probability mass function of X. Recall that a mode of the probability function $p(x)$ is a point x^* for which $p(x^*) \geq p(x)$, for any x. For each $i = 1, \ldots, d$, define the function

   $$m_i(x_i) := \max_{x_j, j \neq i} p(x_1, \ldots, x_d).$$

   Suppose that the maximum of $m_i(x_i)$ is uniquely attained at x_i^*, i.e., $m_i(x_i^*) > m_i(x_i)$, for all $x_i \in \mathcal{X}$. Show that $x^* = (x_1^*, \ldots, x_d^*)$ is the unique mode of p.

   (c) Let $X = (X_1, \ldots, X_d)^T \in \mathbb{R}^d$ be a random vector, with bivariate marginal densities $p_{ij}(x_i, x_j) > 0$ and univariate marginal densities $p_i(x_i)$. Let $G = (V, E)$ be a tree graph on $\{1, \ldots, d\}$, so that G does not contain any cycles. Consider the family of functions

   $$f_m(x_1, \ldots, x_d) = \prod_{i=1}^{d} p_i(x_i)^{m_i} \prod_{(i,j) \in E} p_{ij}(x_i, x_j),$$

   where $m_i \in \mathbb{Z}$ are integers. Find a set of integers $m_1, \ldots, m_d \in \mathbb{Z}$ for which the function f_m is a probability density; i.e., f_m is nonnegative and integrates to one.

2. Exponential families

   (a) Recall that a member of an exponential family of distributions has a density that can be written in the form given in Chapter 8, Eq. (8.3). For the following distributions, verify whether they belong to the exponential family and, if so, write them in the form given in Eq. (8.3); otherwise, explain why they do not belong to the exponential family.

   i. $X \sim \text{Unif}([0, p])$, i.e., X is uniformly distributed on the interval $[0, p]$.

   ii. $Y = \exp(X)$, where $X \sim \mathcal{N}(0, \sigma^2)$.

   iii. X is a continuous random variable on $[0, 1]$ with density

$$f(x; a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}, \quad a, b > 0, \tag{1}$$

   and $\Gamma(x) = \int_0^\infty x^{t-1} \exp(-x) dx$ is the Gamma function.

   (b) Let $X \in \{1, \ldots, m\}$ be a discrete random variable with distribution $p(X = j) = p_j$. The entropy of $p(\cdot)$ is given as $H(p) = -\sum_{j \in \{1,\ldots,m\}} p_j \log p_j$ (assume $0 \log 0 = 0$ here). Let $\{\phi_1, \ldots, \phi_d\}$ denote a collection of functions. Consider the following constrained optimization problem

$$\begin{aligned} \max_{p_1,\ldots,p_m} \quad & H(p) \\ \text{s.t.} \quad & p_j \geq 0 \quad j \in \{1, \ldots, m\} \\ & \sum_{j \in \{1,\ldots,m\}} p_j = 1 \\ & \sum_{j \in \{1,\ldots,m\}} p_j \phi_k(j) = \mu_k, \quad k \in \{1, \ldots, d\}, \end{aligned}$$

   where μ_k are given. Compute the Lagrangian for this constrained optimization problem, and show that the solution belongs to an exponential family of distributions.

3. Density estimation

   (a) Repeat the proof of Theorem 26.18 in the class notes, but use Hoeffding's inequality instead of Bernstein's inequality. You will get a weaker result. Why does Bernstein's inequality yield the correct rate while Hoeffding's inequality does not?

   (b) Show that if there are ties in the data then cross-validation can lead to $\widehat{h} = 0$ in density estimation problem. There are ties in the data if $x_i = x_j$ for some $i \neq j$. Can you suggest a way to fix this problem?

   Hint: consider $d = 1$ and use the boxcar kernel.

   (c) In the class, you have seen how to estimate a density using kernels. There is an even simpler way to estimate a density, using regular histograms. Let $X \in \mathbb{R}$ be a random

variable with unknown density f_X supported on $[0,1]$. The regular histogram estimator partitions the interval $[0,1]$ into D bins of equal length,

$$[0,1] = \cup_{j=1}^{D} \mathrm{Bin}(j) = \bigcup_{j=1}^{D} \left[\frac{j-1}{D}, \frac{j}{D}\right].$$

Denote $B(x) := \mathrm{Bin}(j)$ for which $x \in \mathrm{Bin}(j)$. The estimate $\widehat{f}_{X,D}(x)$, based on n data points $\{X_i\}_{i=1}^n$, can be expressed as

$$\widehat{f}_{X,D}(x) = \frac{D}{n} \sum_{i=1}^{n} \mathbb{I}\{X_i \in B(x)\}.$$

Note that $\widehat{f}_{X,D}(x)$ estimates the density of f_X on the interval $\left[\frac{j-1}{D}, \frac{j}{D}\right]$ by the fraction of points that fall in the $\mathrm{Bin}(j)$.

The only parameter that has to be chosen in the above described procedure is the number of bins D of the regular histogram. The number of bins D can be chosen using the leave-one-out (loo) cross-validation, where the loss

$$L(D) = \int_{[0,1]} \widehat{f}^2_{X,D}(x)dx - 2\int_{[0,1]} \widehat{f}_{X,D}(x) f_X(x)dx$$

is estimated using the following expression

$$\widehat{L}(D) = \int_{[0,1]} \widehat{f}^2_{X,D}(x)dx - \frac{2}{n}\sum_{i=1}^{n} \widehat{f}^{(i)}_{X,D}(X_i), \tag{2}$$

where $\widehat{f}^{(i)}_{X,D}$ is the density estimator obtained after removing the i^{th} observation. Since the loo cross-validation estimator (Eq. (2)) of the loss can be computationally expensive for a large number of samples n, your task is to prove that the following equation

$$\widehat{L}(D) = \frac{2D}{n-1} - \frac{D(n+1)}{n-1}\sum_{j=1}^{D}\left(\frac{|\mathrm{Bin}(j)|}{n}\right)^2 \tag{3}$$

gives a closed form solution for the loo cross-validation estimator. Note that $|\mathrm{Bin}(j)|$ denote the number of observations that fall in the interval $\left[\frac{j-1}{D}, \frac{j}{D}\right]$.

4. Numerical computation

Consider the model $X_{t+1} = aX_t + \epsilon_t$, where $a \in [0,1]$ is a known constant and $\epsilon \sim \mathcal{N}(0, \sigma^2)$ with $\sigma^2 = 0.01$ and $X_0 = 0$. For each of the following values of $a = 0.1, 0.5, 0.95$, draw $n = 500$ independent trajectories of length $T = 100$ from the model. Note that a trajectory is a vector of observation $\mathbf{X}_k = (X_{k,1}, \ldots, X_{k,T})$, $(k = 1, \ldots, n)$. For each of the three values of a, you will estimate the covariance function $\mathrm{cov}(t_1, t_2) = \mathrm{cov}(X_{t_1}, X_{t_2})$, $(1 \le t_1, t_2 \le T)$, using:

(a) `glasso`, which is an `R` implementation of the procedure described in Section 17.7.1 of Chapter 17. You will need to install the R package `glasso`[1].

(b) the thresholding procedure:

$$\widehat{\sigma}_{ij}(M) = \widehat{\sigma}_{ij} \mathbb{1}\left\{ |\widehat{\sigma}_{ij}| \geq M\sqrt{\frac{\log T}{n}} \right\}, \quad i \neq j,$$

where $\widehat{\sigma}_{ij} = \frac{1}{n}\sum_{k=1}^{n}(X_{k,i} - \bar{X}_i)(X_{k,j} - \bar{X}_j)$ is the maximum likelihood estimate of the covariance function $\mathrm{cov}(i, j)$.

You will pick tuning parameters λ for `glasso` and M for the thresholding procedure by maximizing the log-likelihood of the test data. For this assignment, create a test set of $n' = 300$ trajectories of the same length $T = 100$. Summarize your findings.

Finally, give an analytical expression for $\mathrm{cov}(t_1, t_2)$, which should depend only on the difference $|t_1 - t_2|$, the constant a and σ. Compute $\|\widehat{\Sigma} - \Sigma\|_F$, where $\widehat{\Sigma}$ is an estimate of the covariance and Σ consists of elements $\mathrm{cov}(t_1, t_2)$. Which estimation procedure results in a better estimator? Why?

[1] `http://cran.r-project.org/web/packages/glasso/index.html`