Worksheet 2

Problem 1. KL and entropy

The Kullback-Leibler (KL) divergence of a distribution $p(\mathbf{x})$ from another distribution $q(\mathbf{x})$ is given by

$$D_{\mathrm{KL}}(p\|q) = -\int p(\mathbf{x}) \log\left(\frac{q(\mathbf{x})}{p(\mathbf{x})}\right) d\mathbf{x} \,.$$

Prove that $D_{\mathrm{KL}}(p\|q) \geq 0$.

$*$ If you have much time, also think about the following problems. Even if you don't think about them, we will cover them later in this course. These will not be covered in the recitation.

1. The Kullback-Leibler (KL) divergence of a distribution $p(\mathbf{x})$ from another distribution $q(\mathbf{x})$ is given by

   $$D_{\mathrm{KL}}(p\|q) = -\int p(\mathbf{x}) \log\left(\frac{q(\mathbf{x})}{p(\mathbf{x})}\right) d\mathbf{x} \,.$$

   Let $p(\mathbf{x}) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $q(\mathbf{x}) \sim \mathcal{N}(\boldsymbol{m}, \boldsymbol{S})$. Calculate $D_{\mathrm{KL}}(p\|q)$.

2. The entropy of a distribution $p(\mathbf{x})$ is given by

   $$H(p) = -\int p(\mathbf{x}) \log p(\mathbf{x}) \, d\mathbf{x} \,.$$

   Calculate $H(p)$ where $p(\mathbf{x}) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Problem 2. Ridge regression ([HaTF] Ex. 3.29)

Recall that in a ridge regression we minimize $\frac{1}{2}\left\|\boldsymbol{r} - \boldsymbol{X}\boldsymbol{w}\right\|^2 + \frac{1}{2}\lambda\left\|\boldsymbol{w}\right\|^2$. Suppose we run a ridge regression with parameter λ on a single variable x and get coefficient w (so the data matrix $\boldsymbol{X}$ is $N \times 1$, which can be denoted as a vector $\boldsymbol{x} \in \mathbb{R}^N$). We now include an exact copy $x^* = x$ and refit our ridge regression. Show that both coefficients are identical, and derive their value. Show in general that if m copies of a variable x_j are included in a ridge regression, their coefficients are all the same.

Problem 3. Elastic net ([HaTF] Ex. 3.30)

Consider the elastic-net optimization problem:

$$\min_{\boldsymbol{w}} \; \left\|\boldsymbol{r} - \boldsymbol{X}\boldsymbol{w}\right\|^2 + \lambda\left[\alpha\left\|\boldsymbol{w}\right\|^2 + (1-\alpha)\left\|\boldsymbol{w}\right\|_1\right] \,.$$

Show how one can turn this into a lasso problem using an augmented version of $\boldsymbol{X}$ and $\boldsymbol{r}$:

$$\tilde{\boldsymbol{X}} = \begin{bmatrix} \boldsymbol{X} \\ \gamma\boldsymbol{I} \end{bmatrix}$$

and

$$\tilde{\boldsymbol{r}} = \begin{bmatrix} \boldsymbol{r} \\ \mathbf{0} \end{bmatrix} \,.$$

Problem 4. Kernel

1. Suppose $K(\boldsymbol{x}) \geq 0$ and $\int_{\mathbb{R}^d} K(\boldsymbol{x})d\boldsymbol{x} = 1$. Show that the kernel estimator

$$\hat{p}(\boldsymbol{x}) = \frac{1}{Nh^d}\sum_{n=1}^{N} K\left(\frac{\boldsymbol{x}-\boldsymbol{x}_n}{h}\right)$$

   is a density.

2. Suppose $K(\boldsymbol{x}) = \left(\frac{1}{\sqrt{2\pi}}\right)^d \exp\left[-\frac{\|\boldsymbol{x}\|^2}{2}\right]$. Show that each $K\left(\frac{\boldsymbol{x}-\boldsymbol{x}_n}{h}\right)$ can be written as a product of d univariate kernels.

Problem 5. Smoother ([HaTF] Ex.6.8)

Suppose for continuous response Y and predictor X we model the joint density of X,Y using a multivariate Gaussian kernel estimator. This means that

$$\hat{p}(x, y) = \frac{1}{Nh^2}\sum_{n=1}^{N} K_h(x - x_n)K_h(y - y_n)$$

where $K_h(x) = K(x/h)$ and K is the Gaussian kernel. (cf. Problem 4 above.) Show that the conditional mean $\mathbb{E}[Y|X]$ derived from this estimate is a Nadaraya-Watson estimator.

Problem 6. EM (Midterm exam, Fall'21, Problem #-1)

Hilbert owns a PS5, an Xbox and a Switch (three different gaming systems). On each system there is a game. The outcome of each game is either win ("W") or loss ("L"). Every day, he plays the Switch game. If he wins, he continues to play the PS5 game and records the outcome of the PS5 game; otherwise, he continues to play the Xbox game and records the outcome of the Xbox game. The outcomes he recorded for the last ten days of March are as follows:

W W L W W L W W L L

Suppose the event on each day is independent. Denote the (unknown) probabilities of "W" for the PS5, the Xbox and the Switch games by p, q, π, respectively. Let y denote the random variable representing the final outcome, so that $\mathrm{y} = 1$ if "W" is recorded, and $\mathrm{y} = 0$ if "L" is recorded.

1. Suppose we use an expectation-maximization (EM) algorithm to find the maximum-likelihood solution of p, q, π. In plain language, describe what is the latent variable and the values it can take.

2. For the final outcome y, consider its parametric likelihood $p(y \mid p, q, \pi)$. Is it true or false that $p(y \mid p, q, \pi) = \sum_z (p(z \mid p, q, \pi) + p(y \mid z, p, q, \pi))$, where the summation is over all possible values of the latent variable?

3. Write $p(y \mid p, q, \pi)$ as a function of y, p, q, π.

4. Using the data recorded on the last ten days of March, and the initial values

$$(p^{(0)}, q^{(0)}, \pi^{(0)}) = (0.5, 0.5, 0.5),$$

   implement the EM algorithm for one E-step and one M-step. Calculate $(p^{(1)}, q^{(1)}, \pi^{(1)})$.