Foundations of data science, summer 2020

JONATHAN LENNARTZ, MICHAEL NÜSKEN, ANNIKA TARNOWSKI

11. Exercise sheet
Hand in solutions until Thursday, 2 July 2020, 12:00

Exercise 11.1 (Mixture of densities). (8 points)

Suppose you are given some random variables $X^{(i)} \xleftarrow{⚅} \mathbb{R}$ with density p_i. 8
For the computer scientist: some routine `Xi` produces samples of $X^{(i)}$. How do you construct a routine `X` that samples acc. to the overlayed density $\sum_{i<k} w_i p_i$? Prove correctness:

Theorem. *Consider $X^{(i)} \xleftarrow{⚅} p_i$ for $i < k$ and $\hat\imath \xleftarrow{⚅} w$, reading w as a distribution on $\mathbb{N}_{<k}$. Finally, let $X \leftarrow X^{(\hat\imath)}$. Then $X \sim p$.*

Hint: $X \sim p$ means that p is the density of X, ie. $\operatorname{prob}(X \in [a,b]) = \int_a^b p(x)\,\mathrm{d}x$ for all $a < b$.

Remark: This generalizes to random variables with other outputs instead of values in $\mathbb{R}$.

Solution. To show that the density of X is p means that we have to show

$$\operatorname{prob}(X \in [a,b]) = \int_a^b p(x)\,\mathrm{d}x$$

by definition. We have

$$X \in [a,b] \Leftrightarrow \exists i < k \colon X^{(i)} \in [a,b] \wedge w = i.$$

This directly follows from the construction of X. So

$$\begin{aligned}\operatorname{prob}(X \in [a,b]) &= \sum_{i<k} \operatorname{prob}(X \in [a,b] \,|\, \hat\imath = i) \cdot \operatorname{prob}(\hat\imath = i) \\ &= \sum_{i<k} \operatorname{prob}\left(X^{(i)} \in [a,b] \,\middle|\, \hat\imath = i\right) \cdot \operatorname{prob}(\hat\imath = i).\end{aligned}$$

Since $X^{(i)}$ and $\hat\imath$ are independent, we have

$$\operatorname{prob}\left(X^{(i)} \in [a,b] \,\middle|\, \hat\imath = i\right) = \operatorname{prob}\left(X^{(i)} \in [a,b]\right).$$

Thus

$$\begin{aligned}
\operatorname{prob}\left(X \in [a,b]\right) &= \sum_{i<k} \operatorname{prob}\left(X^{(i)} \in [a,b]\right) \cdot w_i \\
&= \sum_{i<k} w_i \cdot \int_a^b p_i(x)\ \mathrm{d}x \\
&= \int_a^b \sum_{i<k} w_i p_i(x)\ \mathrm{d}x \\
&= \int_a^b p(x)\ \mathrm{d}x\,,
\end{aligned}$$

which was what we wanted. ○

Exercise 11.2 (Application of the SVD). (0+13 points)

In this exercise you shall play with the example from

Alex Thomo (2009). Latent Semantic Analysis (Tutorial).

+3 (i) Reprogram it, denote by k the used dimension.

+5 (ii) Examine the resulting ranking if…

(a) …you modify $k \in \{2, 3, 4, 5\}$.

(b) …you omit the scaling step.

(c) …you change the selection of words by omitting words that only occur in a single document or by adding more words.

(d) …you use the Euclidean metric instead of the angle metric.

That's a total of at least 24 cases. You need a careful analysis to isolate important insights.

+5 (iii) Redo similar analysis with a larger dataset: You will find documents `11-document*.txt` in the exercises folder, which contain (parts of) the short overviews of some Wikipedia articles.

Hint: We expect you to present an analysis with insights, explanations and arguments. So, no large tables or thelike.

Solution (Hints). There is no optimal solution to this exercise, because your choices in the implementation and the sheer possibilities in experimenting can lead to a lot of different results, that might not necessarily be always true. The important aspect in a solution are well-planned experiments and a critical treatment of the result.

In general, a higher k leads to more accurate results, but for some queries, even k high did not always lead to the expected solution, as our sample size was quite small. ○