Medical Data Science
SS 19
Tutor:

1	2	3	$\sum$

Michael Graf 3840238
Sabri Aidi 4128027

Assignment 02

(Due May 23rd)

Problem 1

Kernel Methods:

a) Alternative splicing describes the process which makes it possible to translate more proteins than there are protein encoding genes. This is realized by stitching parts of the genome together in different orders, the parts of the genome which are translated, called exons, are stitched together at the donor and acceptor splice site, the sequence between those sites, called intron, is not translated.

b) A similarity is that both kernels match k-mers that are not at exactly the same place in the sequence. A key difference between the spectrum kernel and the WDS kernel is that while the WDS kernel does allow for shifts in position, the matches at shifted positions are weighed down, the spectrum kernel is completely position-independent. Both kernels share the parameter k, defining the maximum length of k-mers that are compared. The WDS kernel has an additional two additional parameters $S(l)$ and δ_s, controlling the maximum shift and the downweighing of matches at shifted positions.

c) p : positions where the Oligomer occurs.
σ : the degree of positional uncertainty, the lower its value the more sharp the gaussian is which means more certainty about the position of Oligomer. The higher its value the smoother the function is which means more uncertainty about the position of occurrence.
This allows to control whether the model should be completely position-independent like the spectrum kernel or position dependent like the weighted degree kernel (without shifts), or anything in between those two extremes.

Domain Adaptation and Multitask Learning:

a) For some Tasks, certain experiments (including data collection) facilitating the analysis of particular processes or phenomena can be performed more readily than for others. This understanding can then be transferred to other tasks, for instance by verifying or refining models of the processes often at a fraction of the original cost.

First Approach :
Train two baseline classifiers one on the source and the other on the target data, then reuse the two generated optimal prediction functions to combine them in a convex manner.

Second Approach :
Re-weight the source labeled samples such that it statistically assimilate the distribution of the target samples. The data can then be used to improve performance when training the target task.

b) Multi-task learning describes the approach of learning multiple functions from and input space to an output space, called tasks, with a tree describing the relationship between the tasks. The objective is to to learn a model for a task while utilizing the data or information from the other available tasks in the tree. Three approaches have been described in the lecture, in which the first two utilize the topology of the tree in the regularizer and the last one uses the topology of the tree directly in the weight vector.

c) MHC are genes that encode proteins found on the surface of cells, which bind antigens that control whether the cells are recognized as pathogens or as foreign cells. Meaning the MHC control the response of the immune system via t helper and t cytotoxic cells.

Problem 2

Since the Kernel $k(x_i, x_j)$ is defined as the inner product of the vectors $\phi(x_i)$ and $\phi(x_j)$ after solving the squared euclidean distance equation the result is:

$$\phi(x_i)^\top \phi(x_i) + \phi(x_j)^\top \phi(x_j) - 2\phi(x_j)^\top \phi(x_i)$$

which can be expressed via the kernel function resulting in the new squared distance equation:

$$k(x_i, x_i) + k(x_j, x_j) - 2k(x_i, x_j)$$

which does not explicitly use the mapping function ϕ but only uses the implicit mapping via the kernel function.

Problem 3

a) see code

b) Heatmap displayed in Figure 1

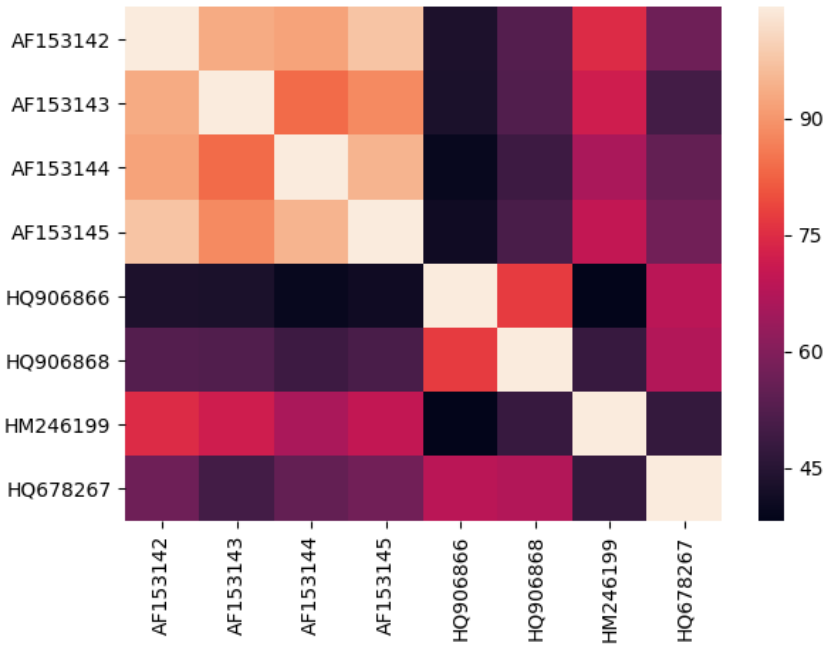

Abbildung 1: Heatmap showing the result of the weighted degree kernel with maximal k-mer length 3