# sheet07a

## Unknown Author

June 1, 2015

## 1 Programming Exercise: Genes (40 P)

In this exercise, various degree kernels including the weighted degree kernel (WDK) will be built for the classification of gene sequences. We will use Scikit-Learn (http://scikit-learn.org/) for training SVMs on various kernels. While Scikit-Learn takes care of the SVM optimization and hyperparameter selection, the focus of this exercise will be the computation of the weighted kernels. The following code is provided to read gene sequences:

```
In [4]:  import numpy
         from sklearn import svm

         # Store data in numpy arrays
         Xtrain = numpy.array([numpy.array(list(l)[:-1]) for l in open(’splice-data/splice-trai
         Xtest  = numpy.array([numpy.array(list(l)[:-1]) for l in open(’splice-data/splice-test
         Ttrain = numpy.array([int(l[:-1]) for l in open(’splice-data/splice-train-label.txt’,’
         Ttest  = numpy.array([int(l[:-1]) for l in open(’splice-data/splice-test-label.txt’,’r
```

### Part A: Degree Kernels (30 P)

The degree kernel of degree $d$ is defined as:

$k_d(x, x') = \sum_{l=1}^{L-d+1} \mathbf{1}_{\{u_{l,d}(x)=u_{l,d}(x')\}}$

where $l$ iterates over the whole genes sequence, and $u_{l,d}(x)$ is a subsequence of string $x$ starting at position $l$ and of length $d$, and $\mathbf{1}_{\{\}}$ is an indicator variable.

**Tasks**:

1. *Implement* a function that computes the kernel matrices for the kernel of degree $d \in \{1, 2, 3, 4\}$.
2. *Run* the code below that outputs the training and test error for each degree kernel.

If your code is efficient, the program below should run in less than 1 minute.

```
In [5]:  import solution

         Ktrains = [None]*4
         Ktests  = [None]*4

         for i in range(4):
             Ktrains[i],Ktests[i] = solution.getdegreekernels(Xtrain,Xtest,i+1)
             mysvm = svm.SVC(kernel=’precomputed’).fit(Ktrains[i],Ttrain)
             Ytrain = mysvm.predict(Ktrains[i])
             Ytest = mysvm.predict(Ktests[i])
             print(’degree: %d   training accuracy: %.3f   test accuracy: %.3f’%(i+1,(Ytrain==T
degree: 1   training accuracy: 0.994   test accuracy: 0.916
degree: 2   training accuracy: 1.000   test accuracy: 0.937
degree: 3   training accuracy: 1.000   test accuracy: 0.964
degree: 4   training accuracy: 1.000   test accuracy: 0.958
```

### Part B: Weighted Degree Kernel (10 P)

We consider a weighted degree kernel with uniform weights:

$k(x, x') = \sum_{d=1}^{4} k_d(x, x')$

where $k_d(x, x')$ is a kernel with degree $d$.

**Tasks:**

1. *Construct* the kernel matrices for the weighted degree kernel.
2. *Compute* the training and test accuracy of an SVM trained with this kernel using the same method as in Part A.

```
In [6]: import solution

        solution.wdk(Ktrains,Ktests,Ttrain,Ttest)
```

```
training accuracy: 1.000   test accuracy: 0.967
```

### Submission guidelines

To facilitate grading, please produce a PDF document from your notebook. This can be done easily by running the following command:

```
ipython nbconvert --to latex sheet07a.ipynb --post PDF
```