Appendix A. Introduction to Binary Support Vector Machines

Below we summarize the main ideas behind binary Support Vector Machines via a short theoretical description followed by an example. For a detailed review of SVMs, refer to [Vapnik1998] and [Burges1998].

1. Linear Support Vector Machines

Given a set of n training points from the m-dimensional space $x_1,...,x_n \in \Re^m$ with positive and negative class labels $y_1,...,y_n \in \{-1,1\}$, linear SVMs solve an optimization problem that seeks a maximum margin classifier (i.e. a hyperplane with the maximum margin width that separates training instances of two classes). This classifier is defined by a subset of training data points called *support vectors*. Then unseen data instances (i.e. samples, which were not used for training) are classified based on which side of the hyperplane they fall into. Mathematically, the separation hyperplane is defined by the equation $x^T w + b = 0$ where $w \in \Re^m$ and b come from the solution of the optimization problem. The decision function is then defined by $f(x) = \text{sgn}(x^T w + b)$.

Consider an example cancer diagnostic problem with two possible outcomes, T-cell acute lympboblastic leukemia (ALL) and B-cell ALL, using expression levels of two genes: c-Myb and SNF2. Given a training set of patients (**Figure A1**), our goal is to build an SVM based diagnostic model. There is obviously an infinite number of hyperplanes (lines in this two-dimensional space) separating data points of two classes. Two possible hyperplanes are shown with dotted lines in **Figure A1**. Linear SVMs provide an "optimal" classifier (bold line) that has the maximum margin width. This classifier is based on 5 support vectors (highlighted with circles) – two ALL T-cell data points and three ALL B-cell data points. The decision function applied to the unseen instance 1 (depicted as a triangle with number 1 inside) will classify it as ALL T-cell because it is above the separation hyperplane. By the same token, the unseen instance 2 (triangle with number 2 inside) will be classified as ALL B-cell since it is below the separation hyperplane.

In practice, linear SVMs may be applied to the datasets that are non-separable (i.e. when it is impossible to come up with a hyperplane separating training instances of two classes without errors). The SVM algorithm can handle this case by using tradeoff (cost) parameter C and penalty parameters (slack variables).

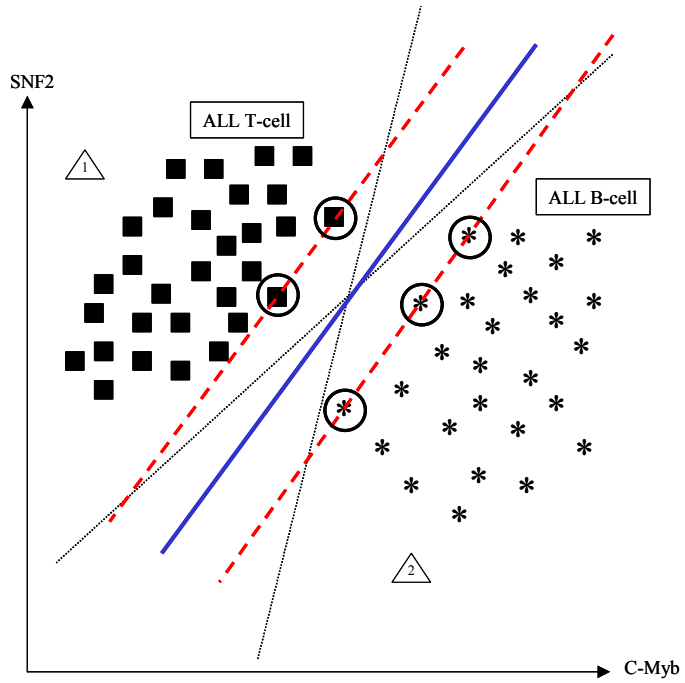

Figure A1. *Binary linear SVMs applied to the example diagnostic problem with two outcomes: ALL T-cell (■) and ALL B-cell (∗).*

2. Non-linear Support Vector Machines

The real power of SVMs comes into play by application of kernel functions, which allow to implicitly map data into a higher dimensional space, called *feature space*. This can be very useful if the training data are non-separable in the input space and become separable in the feature space. Among common kernel functions are linear, polynomial, and radial basis.

Consider an example cancer diagnostic problem involving two possible outcomes, T-cell acute lympboblastic leukemia (ALL) and acute myelogenous leukemia (AML), using expression levels of two genes: c-Myb and SNF2. As it can be seen from the **Figure A2**, data are non-separable in the input space. However, the application of polynomial kernel function maps the data into a higher dimensional feature space where they are separable (**Figure A3**). The resulting hyperplane separating two classes in also shown in the figure. Note that the axes in the feature space are kernel bases dependent on the input variables, expression levels of c-Myb and SNF2.

Similarly to the linear case, slack variables and cost parameter are applicable to non-linear SVMs and can be useful in the situations when the data are still non-separable in the feature space.

It is also important to mention that both linear and non-linear SVMs do not need to directly access to training data points and require only pairwise dot products of the data instances.

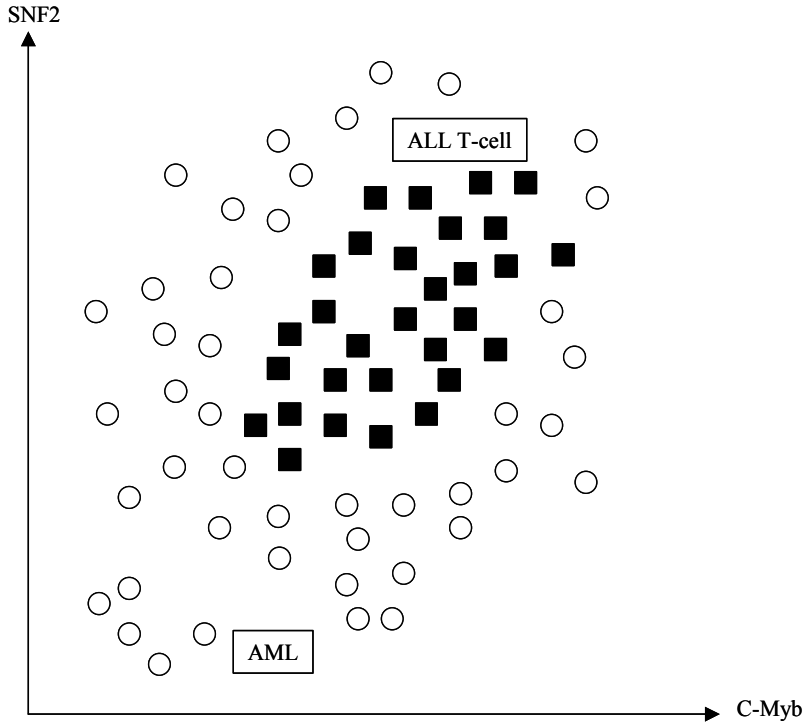

Figure A2. *Input space of the example diagnostic problem with two outcomes: ALL T-cell (■) and AML (○).*

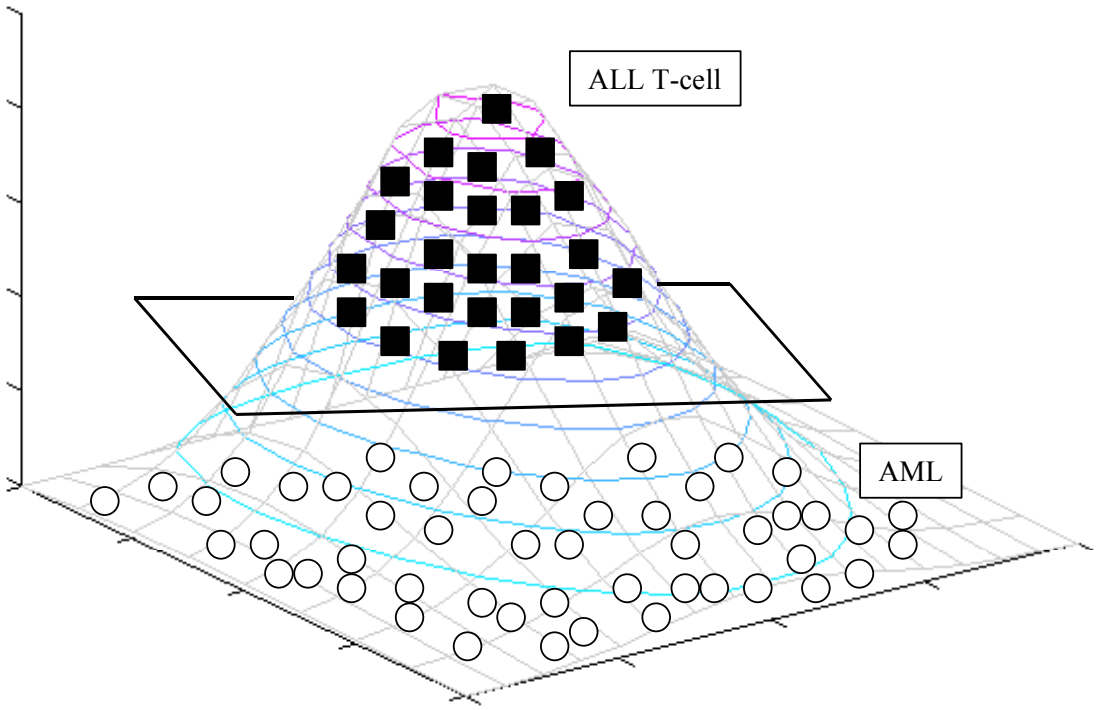

Figure A3. *Feature space of the example diagnostic problem with two outcomes: ALL T-cell (■) and AML (○). The resulting hyperplane is obtained by SVMs with polynomial kernel.*

References

[Vapnik1998] Vapnik, V. "Statistical Learning Theory", Wiley-Interscience, 1998.

[Burges1998] Burges, C. A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery, 2(2):955-974, 1998.