Appendix A

Cell cycle

This appendix contains a summary of the key features of the cell cycle that occurs in dividing, non-germline cells in multi-cellular organisms; mainly summarised from material in Effler et al. (2006) and Alberts et al. (2002), which the reader is encouraged to turn to if they seek more information on the cell cycle. The cell cycle is an important cycle that all dividing cells go through. It consists of four main phases, or states:

1. **G₁ phase.** This is a growth phase, in which the cell grows and builds up materials ready for S phase.

2. **S phase,** in which the new copy of the DNA is synthesised.

3. **G₂,** a second growth phase in which the cell increases in size.

4. **Mitotic phase,** during which mitosis occurs. The M phase consists of two distinct parts: mitosis and cytokinesis. In mitosis the two sets of chromosomes separate, ready for cytokinesis, where the cytoplasm divides in two.

These four phases are illustrated in Figure A.1. There is also a G₀ (resting) phase which the cells can enter into post-mitosis and stay there, or move into G₁ if they commit to replicate further. The majority of mammalian cells are in this phase. Alternatively, cells can go straight from the M phase to G₁. The distinction between G₀ and G₁ is not always clear, and sometimes a combined G₀/G₁ phase is referred to. The M phase is described in more detail below, since this is a key area where genetic changes can influence gross cell abnormalities during cancer. The M phase is divided into two parts, mitosis and cytokinesis, and mitosis can be further subdivided into a number of discrete phases:

1. **Prophase,** in which the chromatids (sets of chromosomes) condense (contract) in a region (of the cell) called the centromere. In most cases of mitosis, the nuclear
Figure A.1. Four main stages of the cell cycle. This figure, from Alberts et al. (2002), shows the four main stages of the cell cycle. These are the two growth phases, G₁ and G₂ in which the cell prepares for the succeeding stages of S (DNA synthesis) and M (mitosis and cytokinesis).

envelope breaks down. An intra-cellular structure known as the spindles form. These govern the separation of the chromatids in the later phases of mitosis.

2. Prometaphase and metaphase, in which the chromosomes congregate at a plane halfway between the two ends. The spindle tapers to these ends. The chromatids attach to this spindle. This plane is called the equatorial plane and is the plane at which the whole cell will divide when nuclear division is complete.

3. Anaphase, in which the two chromatids separate and move to opposite poles, following the line of the spindles.

4. Telophase, where new nuclear envelopes form around the two groups of daughter chromosomes (as they are now called), and the spindles eventually disappear as the cell then divides in two during cytokinesis.

These phases are illustrated in Figure A.2. The cell cycle is a highly regulated process, by a large number of proteins which act to “checkpoint” the cell, and can halt the cell cycle to allow the cell to recover (for example, to repair errors in DNA) or even call for the programmed death (apoptosis) of the cell if the damage is irrecoverable. A number of checkpoints occur. These are:

• One that occurs at the end of G₁ where the cell decides if conditions are favourable to divide. This decision depends on the internal resources of the cell, the external
NOTE: This figure is included on page 179 of the print copy of the thesis held in the University of Adelaide Library.

Figure A.2. Mitotic phase of the cell cycle. This figure, from Alberts et al. (2002), shows the M (mitotic) phase of the cell cycle. This is divided into two main parts: mitosis, and cytokinesis. In mitosis there a number of stages in which the two sets of chromosomes condense and then separate to two poles (ends) of the cell, in preparation for cytokinesis, in which the cell cytoplasm (and set of organelles) divides in two. The cytoskeleton furrows and then cleaves along a line, which eventually forms a bridge between the original parent cell and the new daughter cell. This bridge is then dismantled over a process taking minutes to hours, and the cytoskeletal material is then recycled and used in the next M phase. The interphase period shown is simply the set of other stages: G1, S, and G2, in that order resources available for the next phases, and also signals from other cells on space available and if cell division should be occurring in the location in an organ in which the cell resides.

• Several DNA damage checkpoints, that sense damage:
  1. Before the cell enters S phase (during G1).
  2. During S phase.
  3. After DNA replication (during G2).

• Spindle checkpoints. These detect any problems that have developed during the mitosis part of M phase and before cytokinesis occurs.
B.1 Introduction

This appendix contains an overview of the mathematics behind support vector machines (SVMs). (Binary) SVMs attempt to fit a surface in an $n$-dimensional space, such that it groups the (training) samples of one class on one side, and the samples of another class on the other side. This then gives a function that can classify samples of an unknown class into one of the two classes, hence why this is called a binary SVM. Multiclass SVMs can be built, typically using multiple binary SVMs, and are not discussed here. One would first test the classifier function using samples of known class that are fed into a trained SVM to see if the SVM would perform with a required level of accuracy.

In order to explain SVMs properly, one must first introduce generalised optimisation problems as these are a key component of SVMs, being used to fit surfaces to separate the samples as best as possible. A simple type of SVM is the $C$-SVM, where the $C$ is a classification parameter used in the algorithm. There is a newer type of SVM known as the $\nu$-SVM, which is essentially the same as the $C$-SVM, but with the advantage that the parameter $\nu$ has two useful interpretations as both the upper bound on the fraction of training errors and a lower bound of the fraction of support vectors that would be classified correctly.

B.2 Constrained optimisation problems

SVMs are formulated as a particular constrained optimisation problem. In this section an overview of a generalised constrained optimisation problem is given, based on
B.2 Constrained optimisation problems

Boyd and Vandenberge (2004). A constrained optimisation problem can be written as,

\[
\min_x f_0(x),
\]

subject to,

\[
f_i, h_j : \mathbb{R}^n \rightarrow \mathbb{R}, \quad f_i(x) \leq 0 \forall i, \quad h_j(x) = 0 \forall j,
\]

that is, in general, a mix of inequality constraints \(f_i(x)\), and equality constraints \(h_j(x)\). The primal Lagrangian of the problem is

\[
L_P(x, \lambda, \nu) = f_0(x) + \sum_i \lambda_i f_i(x) + \sum_j \nu_j h_j(x), \quad \lambda_i \geq 0 \forall i.
\]

The dual Lagrangian is then

\[
L_D(\lambda, \nu) = \inf_x L_P(x, \lambda, \nu).
\]

Since the dual Lagrangian is the infimum of the primal Lagrangians, it can become \(-\infty\) for certain values of its arguments. The parameters \(\lambda : \lambda_i \geq 0\) and \(\nu\) for which \(L_D > -\infty\) are called dual feasible. Further, since the dual Lagrangian is a point-wise infimum of a function, so it is concave, even if the primal is not convex.

If we denote the primal optimal point \(p^*\), then because (using Eq. B.3)

\[
\sum_i \lambda_i f_i(x) + \sum_j \nu_j h_j(x) \leq 0,
\]

for a primal feasible point \(x^*\), then

\[
L_D(\lambda, \nu) \leq p^*.
\]

Thus the dual Lagrangian (and hence dual problem) always provides a lower bound to the primal problem. This optimal lower bound can be found by solving the dual problem,

\[
\max_{\lambda, \nu} L(\lambda, \nu)
\]

If we denote the dual optimal point by \(d^*\), the condition \(d^* \leq p^*\) always holds, which is weak duality, strong duality being obtained when \(d^* = p^*\). Strong duality is useful because if it holds, then the dual problem can be solved (which is often easier) and the answer can be converted to the primal domain since we know the solution must
be optimal. In general, strong duality holds if the primal problem is convex and the equality constraints are linear. That is if both of the following conditions hold:

\[ f_i \text{ is convex for } i = 0, \ldots, l , \]

and

\[ h_j(x) = Ax + b. \]

This is true for both types of SVMs that present in this appendix.

As an aside, the primal problem and dual problem can be written more concisely, by noting that

\[
\sup_{\lambda \geq 0, \nu} \inf_{x} \mathcal{L}_P(x, \lambda, \nu) = \begin{cases} 
    f_0(x), & \text{if } x \text{ is feasible}, \\
    \infty, & \text{otherwise}.
\end{cases} \tag{B.8}
\]

Then the primal problem can be written as

\[ p^* = \inf_{x} \sup_{\lambda \geq 0, \nu} \mathcal{L}_P(x, \lambda, \nu), \tag{B.9} \]

and the dual problem as

\[ d^* = \sup_{\lambda \geq 0, \nu} \inf_{x} \mathcal{L}_P(x, \lambda, \nu). \tag{B.10} \]

### B.3 Goal of SVMs

In SVMs, we wish to find a function \( f(x) : \mathbb{R}^n \to \{-1, 1\} \) such that

\[ g(x_i) = y_i = \begin{cases} 
    -1, & x_i \in \mathcal{S}_1, \\
    1, & x_i \in \mathcal{S}_2, \end{cases} \tag{B.11}
\]

where \( \mathcal{S}_1 \) and \( \mathcal{S}_2 \) are two classes that the samples are from. We would then prefer that this function is reasonably accurate (in terms of low numbers of false positives and false negatives). Depending on the application, it may be that we desire a low number of false negatives (since we might want to detect all cancers) but that we are less concerned about false positives (removing a benign lump only has body image implications). The two types of SVM considered are detailed below.
The main algorithm tried in order to train a support vector machine is the C-support vector classification (Cortes and Vapnik 1995, Vapnik 1998). This solves the problem,
\[
\min_{w,b,\xi} \frac{1}{2} w^T w + C \sum_{i=1}^{l} \xi_i, \quad (B.12)
\]
where \(x_i \in \mathbb{R}^n\) subject to
\[
y_i \left( w^T \phi(x_i) + b \right) \geq 1 - \xi_i, \\
\xi_i \geq 0, \\
i = 1, \ldots, l.
\]

The dual problem is
\[
\min_{\alpha} \frac{1}{2} \alpha^T Q \alpha - e, \quad 0 \leq \alpha_i \leq C, \ i = 1, \ldots, l, \quad (B.13)
\]
subject to
\[
y^T \alpha = 0,
\]
where \(e\) is a vector with 1s in all its entries, \(C > 0\) is the upper bound, \(Q\) is an \((l \times l)\) positive semidefinite matrix \(Q_{ij} = y_i y_j K(x_i, x_j)\) for a kernel \(K(x_i, x_j) \equiv \phi(x_i)^T \phi(x_j)\) (see below for more on kernel functions). Once the system is trained, this results in a particular set of parameters \(y_i, \alpha_i, b, i = 1, \ldots, l\). These are then used in a decision function:
\[
g(x) = \text{sgn} \left( \sum_{i=1}^{l} y_i \alpha_i K(x_i, x) + b \right), \quad (B.14)
\]
where \(x\) is the test/unknown sample vector \(x \notin \{x_1, \ldots, x_i, \ldots, x_l\}\) to be classified. The set \(\{x_1, \ldots, x_i, \ldots, x_l\}\) is the set of training samples.

### B.5 \(\nu\)-SVMs

\(\nu\)-SVMs (Schölkopf et al. 2000) uses a different parameter than \(C\), denoted \(\nu \in (0, 1]\) which is an upper bound on the fraction of training errors and a lower bound of the fraction of support vectors.

The primal optimisation problem is
\[
\min_{w,b,\xi,\rho} \frac{1}{2} w^T w - \nu \rho + \frac{1}{l} \sum_{i=1}^{l} \xi_i, \quad (B.15)
\]
subject to

\[ y_i \left( w^T \phi(x_i) + b \right) \geq \rho - \zeta_i, \]
\[ \zeta_i \geq 0, \]
\[ i = 1, \ldots, l, \]
\[ \rho \geq 0 \]

The dual is:

\[ \min_{\alpha} \frac{1}{2} \alpha^T Q \alpha, \]  \hspace{1cm} (B.16)

where \( Q \) is a matrix as defined above for C-SVMs. subject to,

\[ 0 \leq \alpha_i \leq \frac{1}{T}, \]
\[ i = 1, \ldots, l, \]
\[ e^T \alpha \geq \nu, \]
\[ y^T \alpha = 0. \]

The decision function is the same as for C-SVMs.

**B.6 Kernels**

Four kernels can be used in the software package we used, and these are the most common choice of kernels:

1. linear: \( K(x_i, x_j) = x_i^T x_j \),
2. polynomial: \( K(x_i, x_j) = (\gamma x_i^T x_j + r)^d \), \( \gamma > 0 \),
3. radial basis function (RBF): \( K(x_i, x_j) = \exp \left( -\gamma \| x_i - x_j \| \right) \), \( \gamma > 0 \), and
4. sigmoid: \( K(x_i, x_j) = \tanh \left( \gamma x_i^T x_j + r \right) \).

Note that \( \gamma \) is another parameter used in both models. A range of \( (C, \gamma) \in \mathbb{R}^2 \) or \( (\nu, \gamma) \in \mathbb{R}^2 \) values are typically tried for the nonlinear kernels. If a linear kernel is successful, it suggests the data points are linear separable, which means when plotted in \( \mathbb{R}^n \), they fall into two distinct clusters, and hence we would also expect clustering algorithms to cluster the data well.
Bibliography


Bibliography


Folkman-J. (1990). What is the evidence that tumors are angiogenesis dependent?, *Journal of the National Cancer Institute*, 82, pp. 4–6.


Bibliography


**Bibliography**


Biography

Matthew Berryman received a B.Sc. in Mathematical and Computer Sciences and a B.E. (Hons.) in Computer Systems from The University of Adelaide, Australia. In his Ph.D. candidature at The University of Adelaide at the Centre for Biomedical Engineering, he worked on several areas in complex systems including signal processing of DNA and analysis of EEG signals. In 2003 he won a Santa Fe Institute CSSS scholarship and was also a visiting scholar working with Cosma Shalizi at the Center for the Study of Complex Systems at the University of Michigan, Ann Arbor. In 2004 he returned to spend another two and a half months working with Cosma, and in 2005 spent two months working with Neil Johnson at Oxford University. He was a research assistant at the University of South Australia at the Centre for Pharmaceutical Research, applying machine learning techniques to biomarkers and drug discovery. He was the treasurer of the Electrical and Electronic Engineering Society of Adelaide University (EEESAU), an IEEE affiliated student group. Prizes won include the Seimens Prize for Innovation (South Australian winner) in 2003, and the University of Adelaide Research Abroad Scholarship and Society for Mathematical Biology Landahl Scholarship in 2004.
Disclaimer

NOTE: This cartoon strip is included on page 203 of the print copy of the thesis held in the University of Adelaide Library.

Figure B.1. Disclaimer. Used with permission from "Piled Higher and Deeper" by Jorge Cham
www.phdcomics.com