# Advanced Algorithms Course. Lecture Notes. Part 6

## A Machine Learning Problem: Image Segmentation

Our aim is to label every pixel of a digital image as foreground (part of an object) or background. The image is represented as an undirected graph $G = (V, E)$ where nodes are pixels and edges exist between any two neighbored pixels (according to some definition of neighbors). For every pixel $i$ we are also given two numbers $a_i$ and $b_i$ expressing the strength of belief that pixel $i$ is foreground or background, respectively. We do not discuss here in depth how these values are obtained (criteria could be, for example, the colors and positions of pixels). A further assumption is that the picture does not comprise too many switches between foreground and background, that is, it shows a few large and connected objects. Therefore we introduce penalties for label switches: For each pair of neighbored pixels $i, j$ we charge a penalty $p_{ij}$ if $i$ and $j$ have different labels. Altogether this gives rise to the following optimization problem: Split $V$ into sets $A$ and $B$ (foreground and background) so as to maximize

$$q(A, B) := \sum_{i \in A} a_i + \sum_{j \in B} b_j - \sum_{(i,j) \in E, i \in A, j \in B} p_{ij}.$$

That is, the segmentation should respect the classification criteria for the single pixels, but at the same time it should not need too many switches.

We can reduce this problem to Minimum Cut as follows. First observe that the problem is equivalent to minimizing

$$q'(A, B) := \sum_{i \in A} b_i + \sum_{j \in B} a_j + \sum_{(i,j) \in E, i \in A, j \in B} p_{ij}.$$

That is, we want to minimize the penalties for both false labels and switches. As we need a directed graph, we replace every edge $(i, j)$ with two opposite

directed edges with capacity $p_{ij}$. We insert a source $s$ and sink $t$, and for every pixel $k$ we insert edges $(s,k)$ with capacity $a_k$, and $(k,t)$ with capacity $b_k$. Now any $s-t$ cut $(A,B)$ has capacity $q'(A,B)$. Thus, an optimal segmentation corresponds to a minimum cut.

### Project Selection

Let $P$ be a set of possible projects to choose from. Project $i$ has revenue $p_i$. A value $p_i$ can also be negative, in which case the project is an investment for other projects: Some projects depend on others. These dependencies are given as a directed graph $G=(P,E)$ where an edge $(i,j)$ means: if $i$ shall be done, then $j$ must be done, too (before $i$ can even start). Clearly, $G$ must be acyclic, since projects in a directed cycle of dependencies can never be done. We call a set of projects $A \subset P$ feasible if $A$ respects these precedence constraints. The problem is to select a feasible set $A$ that maximizes $\sum_{i\in A} p_i$. This is also known as the Open-Pit Mining problem; one can easily imagine the reason.

We reduce Project Selection (Open-Pit Mining) to Minimum Cut. We insert a source $s$ and a sink $t$. Edges are $(s,i)$ with capacity $p_i$, if $p_i > 0$, and $(i,t)$ with capacity $-p_i$, if $p_i < 0$. Edges in $G$ (for the precedence constraints) get a huge capacity. Hence none of these edges can go from $A$ to $B$ in a minimum cut $(A\cup\{s\}, B\cup\{t\})$. It follows that $A$ is feasible whenever $(A\cup\{s\}, B\cup\{t\})$ is a minimum cut. Now we can solve the Minimum Cut problem and need not worry about the feasibilty of $A$.

It remains to show that minimizing the cut capacity is in fact equivalent to maximizing the revenue. This is proved in a few lines:

$$c(A\cup\{s\}, B\cup\{t\}) = \sum_{p_i>0, i\in B} p_i - \sum_{p_i<0, i\in A} p_i$$

holds by the definition of capacity. We artificially add zero:

$$c(A\cup\{s\}, B\cup\{t\}) = \sum_{p_i>0, i\in B} p_i - \sum_{p_i<0, i\in A} p_i - \sum_{p_i>0, i\in A} p_i + \sum_{p_i>0, i\in A} p_i.$$

Now we can group the terms in a different way:

$$c(A\cup\{s\}, B\cup\{t\}) = \sum_{p_i>0} p_i - \sum_{i\in A} p_i.$$

Note that the first term is constant and the second term is the revenue.

# Randomized Algorithms

## Basics of Probability Theory

This section is not a full-fledged introduction to probability theory, but only a recap of the absolute minimum knowledge needed for a real understanding of randomized algorithms and their analysis.

The mathematical essence of the notion of probability can be described by **Kolmogorov's axioms**, without recurring to any interpretation of proabilities. We deal with a **probability space** which is a set $\Omega$ with a probability function. For simplicity we focus on discrete (finite or countably infinite) sets $\Omega$, which is the most relevant case in algorithmic contexts. Subsets of $\Omega$ are called **events**. The probability $Pr(A)$ of an event $A$ is a number from the interval $[0, 1]$, and probabilities have to satisfy the following simple properties (these are Kolmogorov's axioms): $Pr(\emptyset) = 0$; $Pr(\Omega) = 1$; if $A \cap B = \emptyset$ then $Pr(A \cup B) = Pr(A) + Pr(B)$. The last property called additivity must also hold for countably infinite sets of disjoint events, but this does not matter for finite $\Omega$.

Single-element events $A = \{\omega\}$ are also called elementary events. We may simply write $Pr(\omega)$ instead of $Pr(A) = Pr(\{\omega\})$.

From the axioms it follows immediately that $Pr(\Omega \setminus A) = 1 - Pr(A)$, and $Pr(A \cup B) \leq Pr(A) + Pr(B)$ for any events $A$ and $B$. The latter inequality is so useful that it deserves an own name: we call it the **union bound**. One can use it to bound the probablity of a complicated event which is, however, the disjunction of simpler events with easily computable probabilities.

Sometimes we know already that some event $B$ occurs, and we want to know the probability of $A$, given this additional knowledge. This **conditional probability** is $Pr(A|B) := Pr(A \cap B)/Pr(B)$. Pronounce $Pr(A|B)$ as "probability of $A$ given $B$" or "probability of $A$ conditional on $B$". We call an event $A$ **independent** of an event $B$ if $Pr(A|B) = Pr(A)$. In that case we obviously get $Pr(A \cap B) = Pr(A)Pr(B)$, hence the independence relation is symmetric, and we can simply say "$A$ and $B$ are independent". It is not always intuitive whether two events are independent; then we have to check independence using the definition. Also, do not confuse independent and disjoint events ($A \cap B = \emptyset$) – these are totally different things!

A **random variable** is a function $X$ from a probability space into, e.g., the real numbers. (We only consider the case of real-valued $X$ and discrete $\Omega$.) Formally: $X : \Omega \longrightarrow R$. Every possible value $x$ of $X$ gets a probability

in an obvious way: $Pr(X = x) = Pr(X(\omega) = x)$. The **distribution** of $X$ is $Pr(X = x)$ viewed as a function of $x$. Note that a random variable and its distribution are two different objects. Two random variables with equal distributions are not necessarily the same function on $\Omega$. This distinction is important when we combine several random variables by algebraic operations (see below).

The **expected value** or **expectation** of a random variable $X$ is defined as $E[X] := \sum_{\omega\in\Omega} Pr(\omega)X(\omega)$. Note that $E[X] = \sum_x Pr(X = x) \cdot x$, that is, the expectation depends only on the distribution of $X$. Intuitively, $E[X]$ is the long-term average of $X$ when we observe the random variable many times independently.

A frequent misunderstanding is that $Pr(X > E[X]) = 1/2$, or similar. This is far from being true in general. For instance, let $X$ be the random variable that describes a win in a lottery (where the stake is not considered in $X$). The expected win is some (small) positive amount, but the probability of winning anything is very small, certainly not $1/2$. A "probability-free" formulation of this insight is: The average of a set of values is in general distinct from the median!

Random variables $X$ and $Y$ on the same probability space are called **independent** if $Pr(X = x, Y = y) = Pr(X = x)Pr(Y = y)$ for all values $x$ and $y$. In the same way as for random events we could instead define independence by the property that knowing the value of $X$ has no impact on the distribution of $Y$, and then this "product rule" comes out.

Random variables, without loss of generality defined on the same probability space, can be combined by arbitrary algebraic operations: We simply apply the operation to their random values. For instance, the sum $X+Y$ of random variables $X$ and $Y$ is given by $(X+Y)(\omega) = X(\omega)+Y(\omega)$. Similarly we can define the product, and so on.

A useful and powerful property is the **linearity of expectation**. It says that $E$ is a linear operator, that means, $E[X + Y] = E[X] + E[Y]$. Note that this holds for arbitrary random variables, not only for independent ones. The proof is a straighforward calculation:

$$E[X+Y] = \sum_{\omega\in\Omega} Pr(\omega)(X+Y)(\omega) = \sum_{\omega\in\Omega} Pr(\omega)(X(\omega)+Y(\omega))$$

$$= \sum_{\omega\in\Omega} Pr(\omega)X(\omega) + \sum_{\omega\in\Omega} Pr(\omega)Y(\omega) = E[X] + E[Y].$$

A similar property for the product does not hold in general. We have $E[XY] = E[X]E[Y]$ in special cases only. The most important sufficient condition is that $X$ and $Y$ are independent. Again, the proof is a straighforward calculation, but this time it is easier to work on the range of values rather than on $\Omega$. Also note carefully in which step independence is used:

$$E[XY] = \sum_z Pr(XY = z)z = \sum_z \sum_{x,y:\, xy=z} Pr(X = x, Y = y)xy$$

$$= \sum_z \sum_{x,y:\, xy=z} Pr(X = x)xPr(Y = y)y = \sum_{x,y} Pr(X = x)xPr(Y = y)y$$

$$= \sum_x Pr(X = x)x + \sum_y Pr(Y = y)y = E[X]E[Y].$$