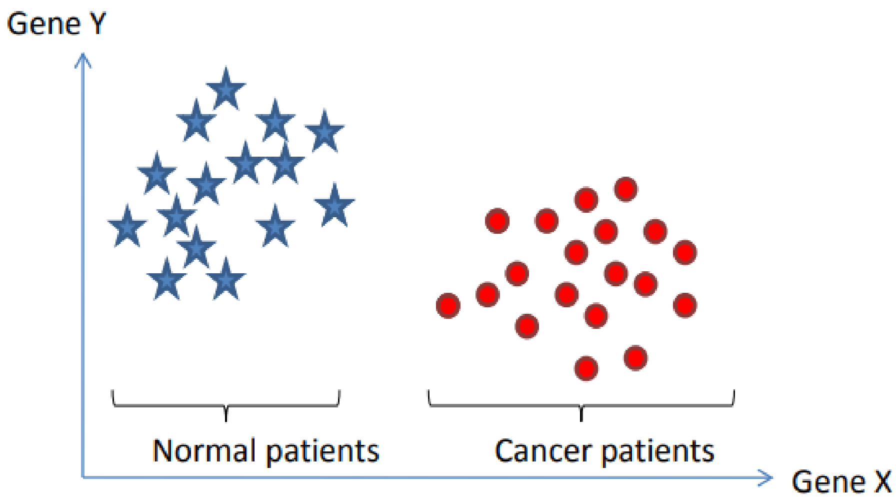
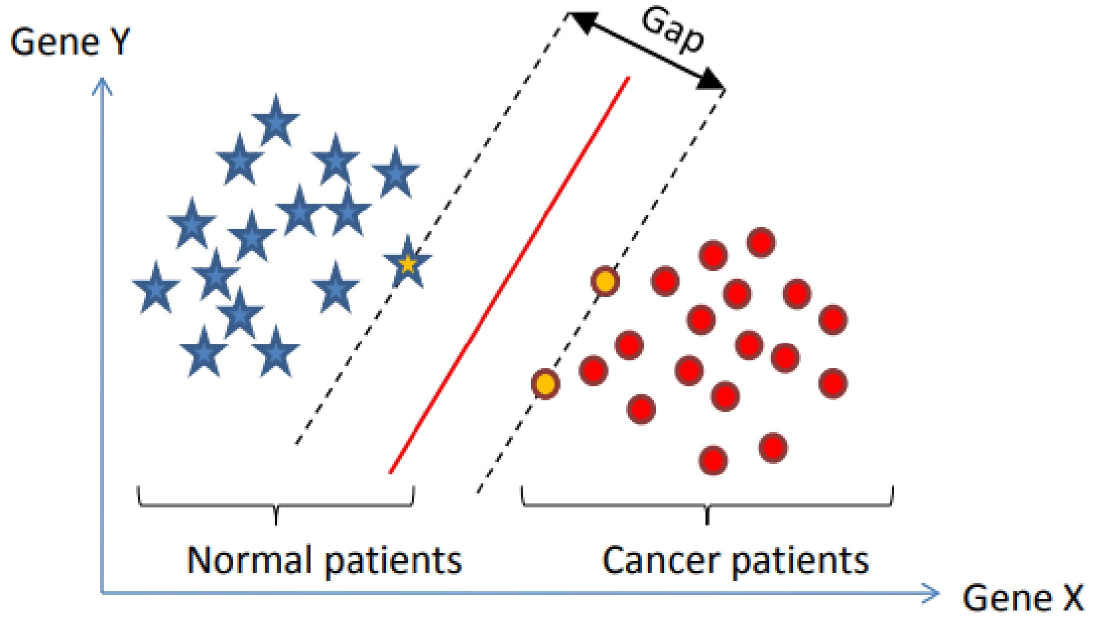
Brief Overview of SVM (Source: https://med.nyu.edu/chibi/sites/default/files/chibi/Final.pdf)

Main ideas of SVMs

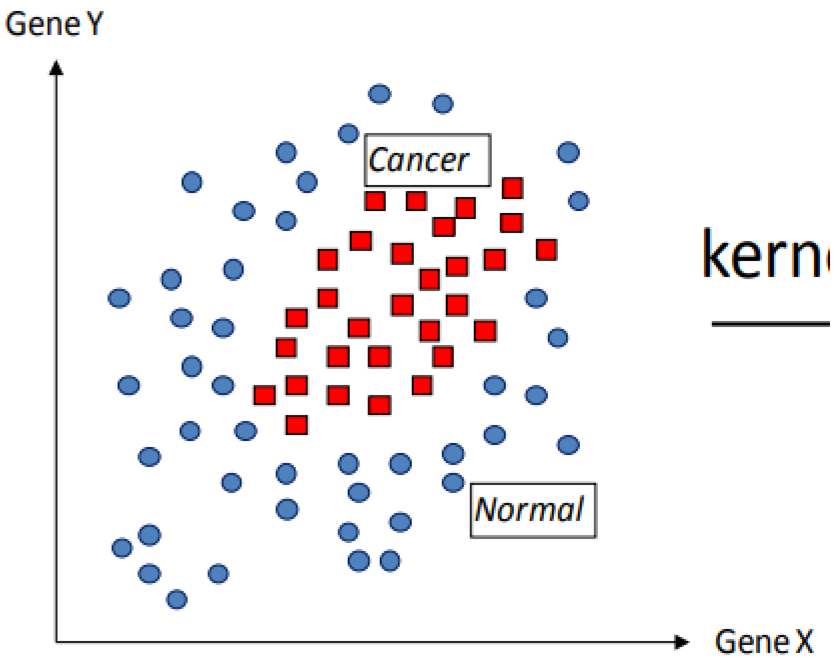


- Consider example dataset described by 2 genes, gene X and gene Y
- Represent patients geometrically (by "vectors")

Main ideas of SVMs

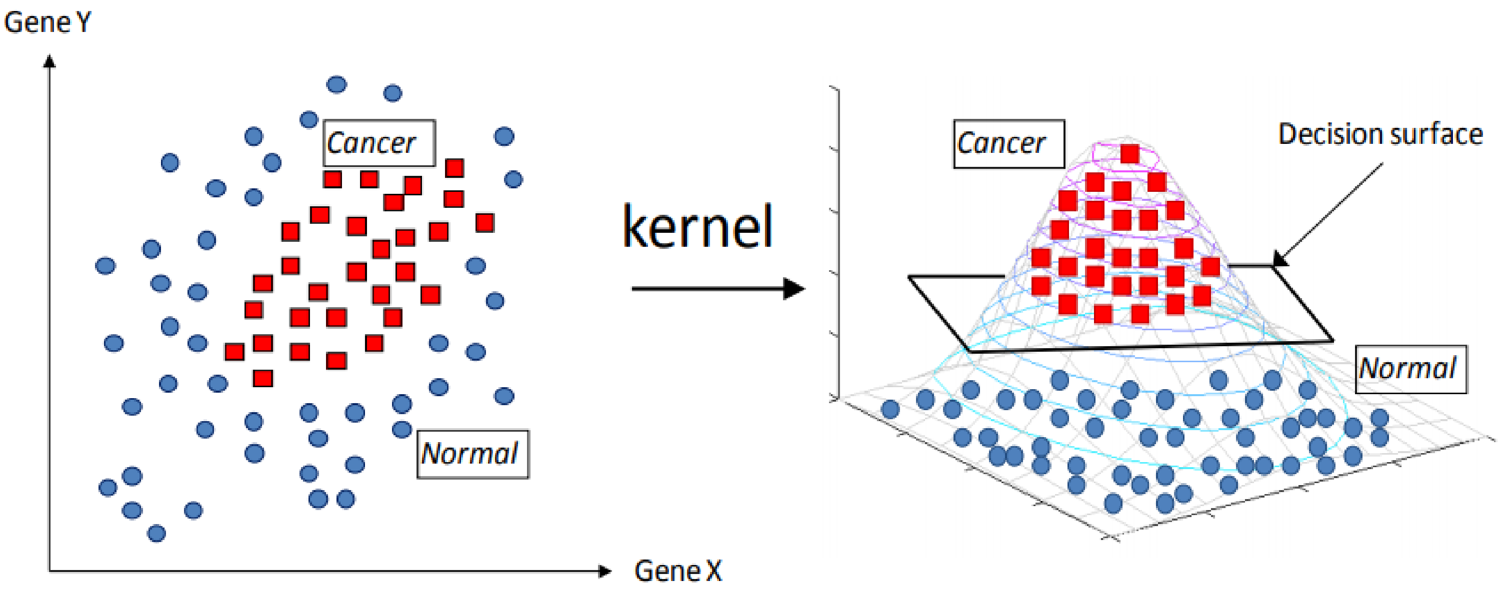


- Find a linear decision surface ("hyperplane") that can separate patient classes <u>and</u> has the largest distance (i.e., largest "gap" or "margin") between border-line patients (i.e., "support vectors");



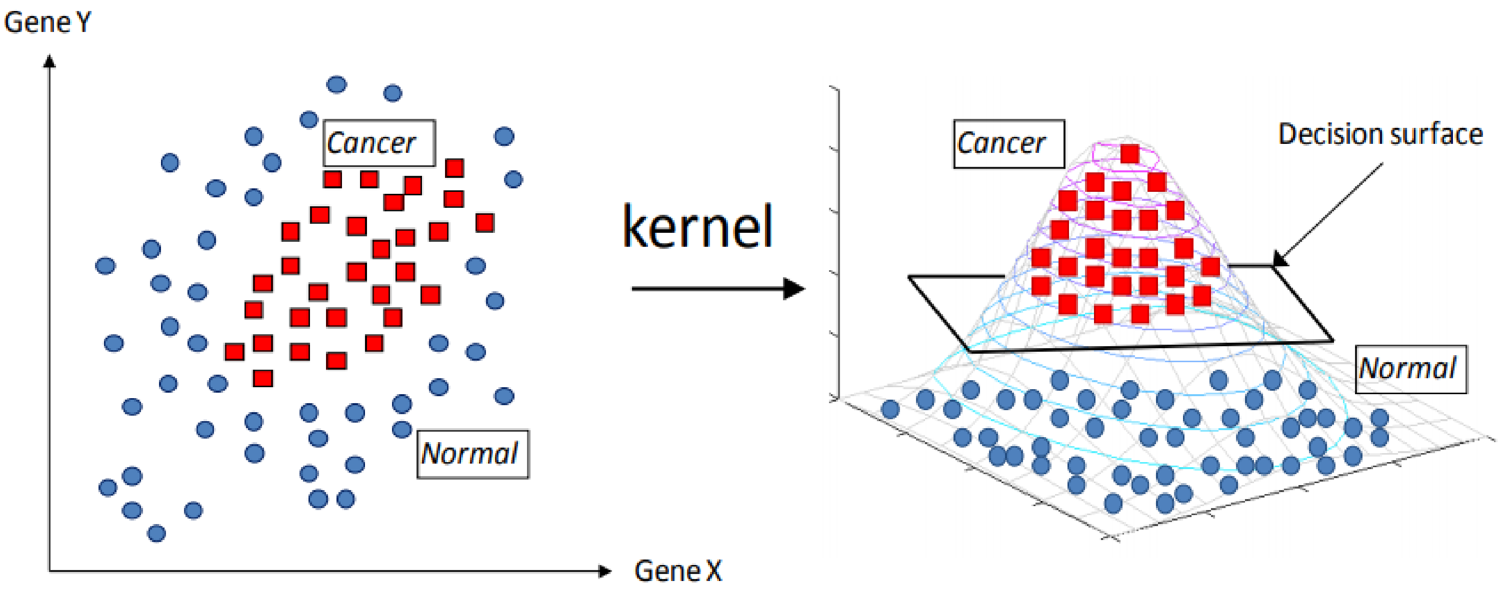
kern

Main ideas of SVMs



- If such linear decision surface does not exist, the data is mapped into a much higher dimensional space ("feature space") where the separating decision surface is found;
- The feature space is constructed via very clever mathematical projection ("kernel trick").

Main ideas of SVMs

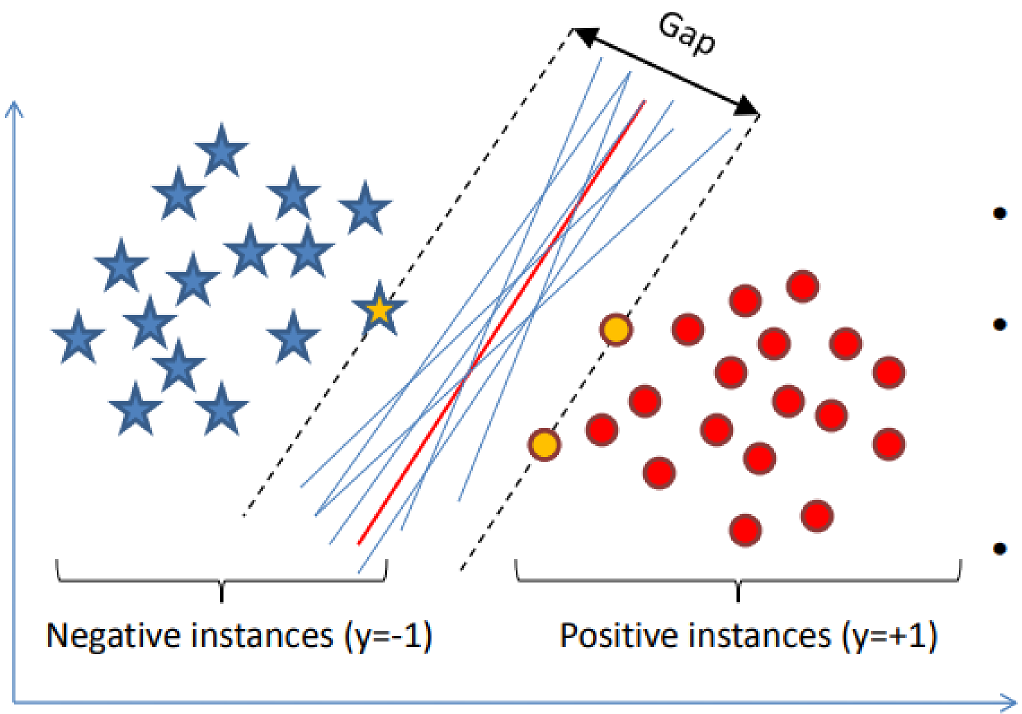


- If such linear decision surface does not exist, the data is mapped into a much higher dimensional space ("feature space") where the separating decision surface is found;
- The feature space is constructed via very clever mathematical projection ("kernel trick").

Case 1: Linearly separable data; "Hard-margin" linear SVM

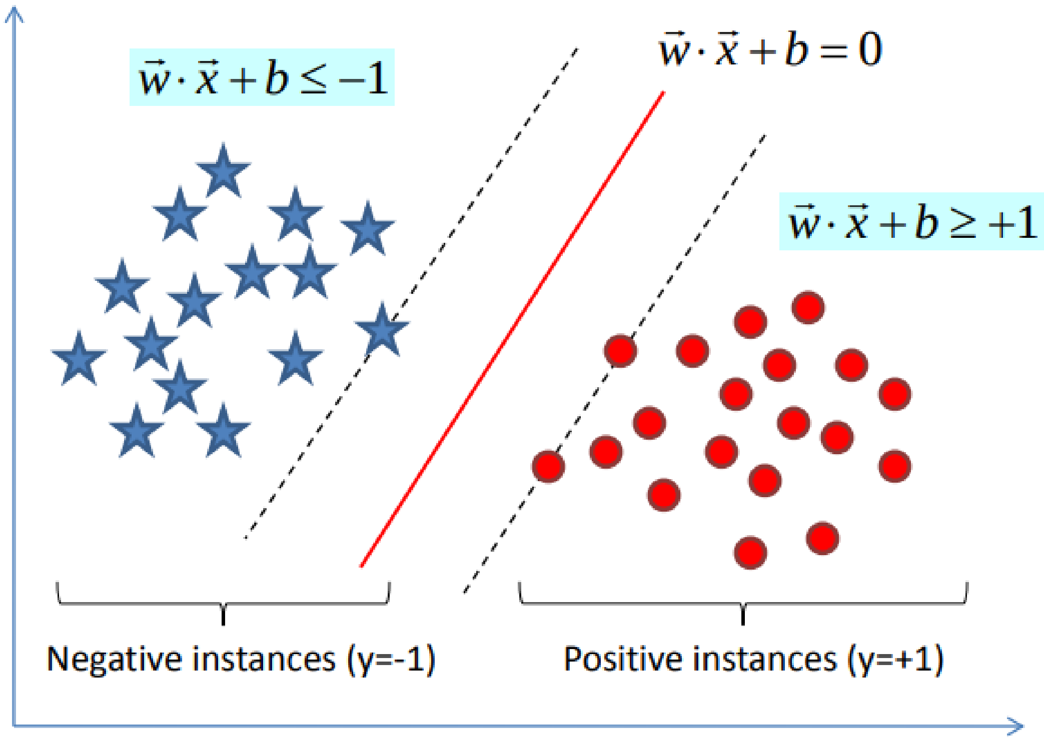
Given training data: $\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_N \in R^n$

$y_1, y_2, \ldots, y_N \in \{-1, +1\}$



- Want to find a classifier (hyperplane) to separate negative instances from the positive ones.
- An infinite number of such hyperplanes exist.
- SVMs finds the hyperplane that maximizes the gap between data points on the boundaries (so-called "support vectors").
- If the points on the boundaries are not informative (e.g., due to noise), SVMs will not do well.

Statement of linear SVM classifier



In addition we need to impose constraints that all instances are correctly classified. In our case:

$$\vec{w}\cdot\vec{x}_i+b\leq -1 \text{ if } y_i=-1$$
$$\vec{w}\cdot\vec{x}_i+b\geq +1 \text{ if } y_i=+1$$

Equivalently:

$$\boxed{y_i(\vec{w}\cdot\vec{x}_i+b)\geq 1}$$

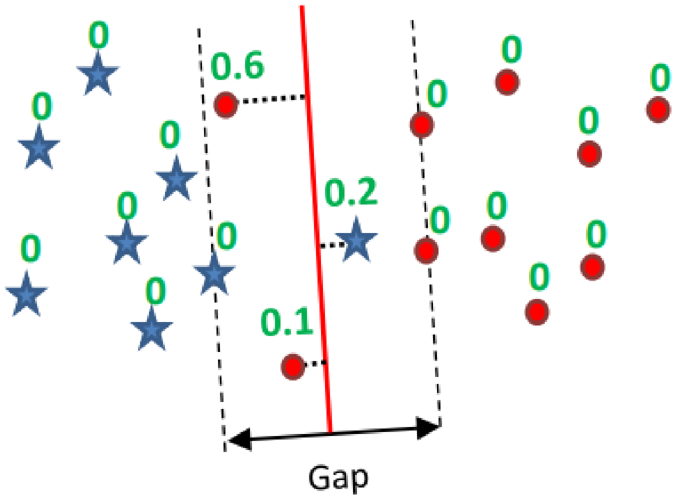
<u>In summary:</u>

Want to minimize $\frac{1}{2}\left\|\vec{w}\right\|^2$ subject to $y_i(\vec{w}\cdot\vec{x}_i+b)\geq 1$ for $i=1,\ldots,N$

Then given a new instance x, the classifier is $f(\vec{x})=sign(\vec{w}\cdot\vec{x}+b)$

Case 2: Not linearly separable data; "Soft-margin" linear SVM

What if the data is not linearly separable? E.g., there are outliers or noisy measurements, or the data is slightly non-linear.



Want to handle this case without changing the family of decision functions.

Approach:

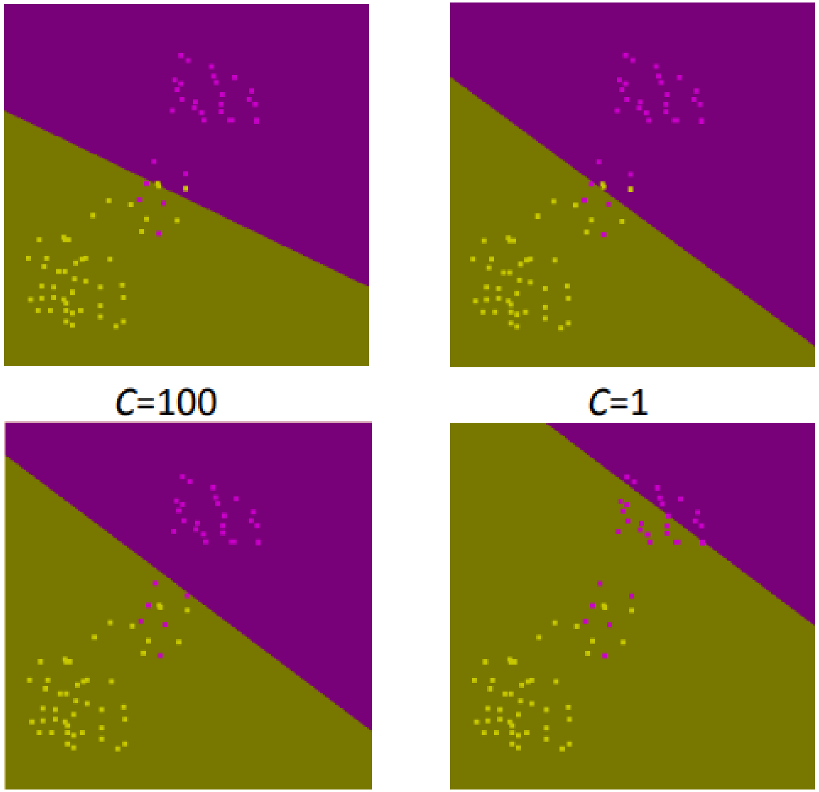
Assign a "slack variable" to each instance $\xi_i \geq 0$, which can be thought of distance from the separating hyperplane if an instance is misclassified and 0 otherwise.

Want to minimize $\frac{1}{2}\|\vec{w}\|^2 + C\sum_{i=1}^{N}\xi_i$ subject to $y_i(\vec{w}\cdot\vec{x}_i + b) \geq 1 - \xi_i$ for $i = 1,\ldots,N$

Then given a new instance x, the classifier is $f(x) = sign(\vec{w}\cdot\vec{x} + b)$

Parameter C in soft-margin SVM

Minimize $\frac{1}{2}\left\|\vec{w}\right\|^2 + \boxed{C}\sum_{i=1}^{N}\xi_i$ subject to $y_i(\vec{w}\cdot\vec{x}_i + b) \geq 1 - \xi_i$ for $i = 1,\ldots,N$



C=100 C=1

C=0.15 C=0.1

- When C is very large, the soft-margin SVM is equivalent to hard-margin SVM;
- When C is very small, we admit misclassifications in the training data at the expense of having w-vector with small norm;
- C has to be selected for the distribution at hand as it will be discussed later in this tutorial.