Support Vector Machines

Outline

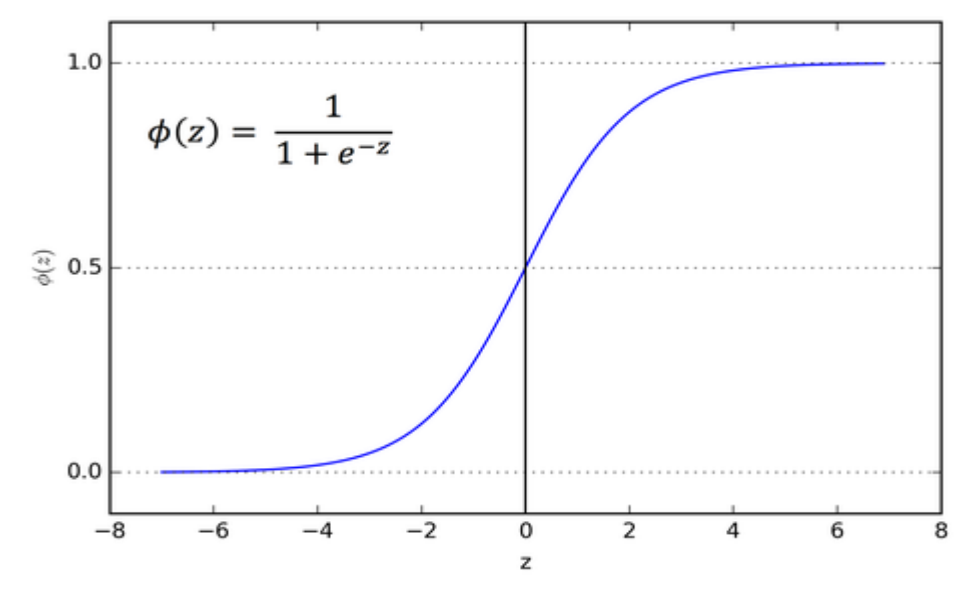
- Concept of Large Margin Classification
  - Functional
  - Geometric
- The SVM large margin classifier
  - Functional
  - Geometric
- The Optimization Function – optimization process itself will not be discussed
- Concept of Kernels for non-linear classification.

Basic Introduction

- Original SVM invented by Vladimir Vapnik and co-workers in Russia in 1963
- Nonlinear SVM developed in 1995 by Vapnik & others
- Immense Applications of SVMs
  - In Image Classification & Segmentation
  - In Natural Language, particularly text, processing
  - Wide applications in Biological Sciences
- Was highest in accuracy among all ML algorithms till the third wave of ANN's after 2010 – served as a comparison benchmark in many developments.

Concept of Large Margin Classification

- Let us revisit the Sigmoid function



- Recall our hypothesis in supervised learning:

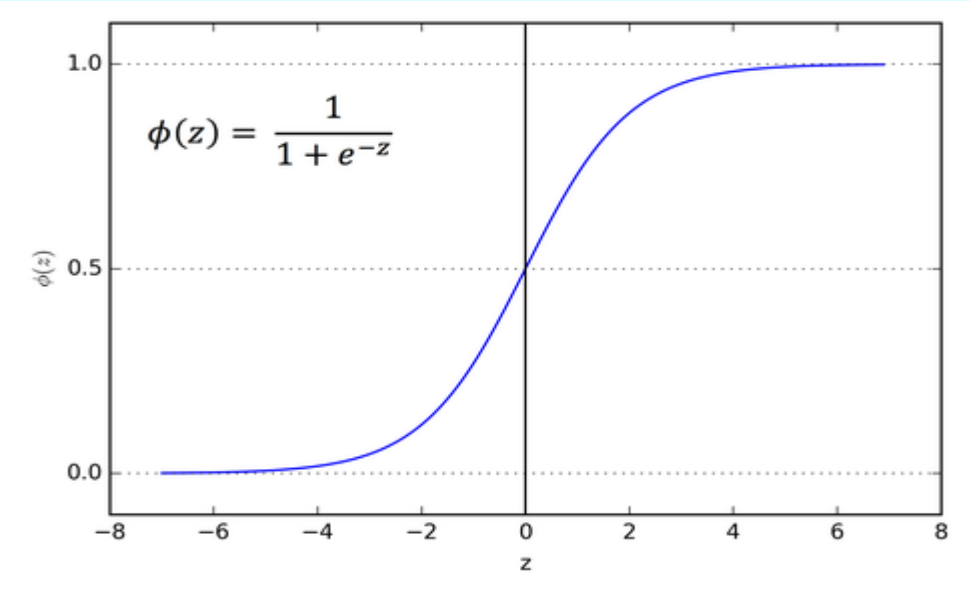
$$h_{\theta}(\mathbf{x}) = g(\boldsymbol{\theta}^T \mathbf{x}) = g(z)$$

in the illustration $\phi(.)$ is shown instead of g(.)

  - note that θ and x are both vectors and shown in bold in the above expression; down the line we will keep them in normal font but just be aware that these two are vectors.

Concept of Large Margin Classification

- As in typical supervised learning, let us assume that we have a training data set (x^i, y^i), i = 1,…, m, where y^i is either +1 or 0
- Further, refer to sigmoid function plot below, we know that if $\theta^T x^i > 0$, then $g(\theta^T x^i) > 0.5 \Rightarrow h_\theta(x^i) > 0.5 \Rightarrow$ we classify data point *i* as +1
- Then can we not say that if for some *i*, $\theta^T x^i >> 0$, then $g(\theta^T x^i) >> 0.5 \Rightarrow h_\theta(x^i) >> 0.5 \Rightarrow$ we can classify that point as +1, more confidently and correctly?
- And exactly analogous logic holds for cases of $\theta^T x^i < 0$?

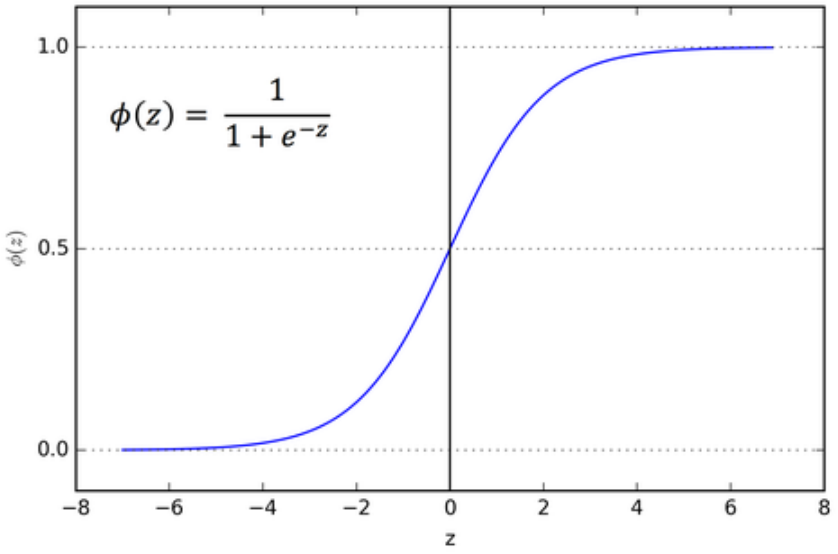


Concept of Large Margin Classification

- Putting it more methodically, we can say that

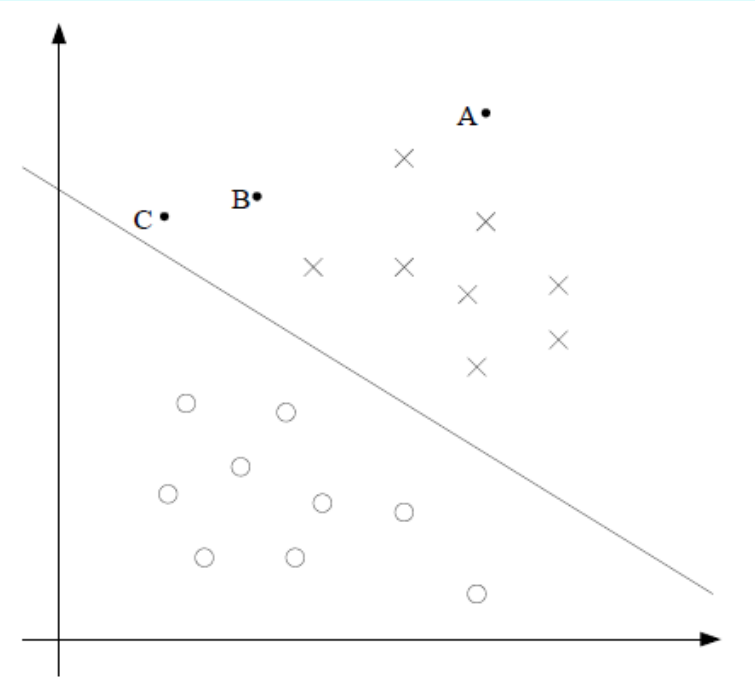
$\theta^T x$	$g(\theta^T x)$	Prediction	Remarks
> 0	> 0.5	$y^* = 1$	**The more strongly $\theta^T x$ > or < 0, the closer is $g(\theta^T x)$ to +1 or 0, and the more confident, reliable and correct our prediction.**
< 0	< 0.5	$y^* = 0$	

- Hence, the farther $\theta^T x$ is from the boundary (*0 in this case*), the more reliable the prediction!
- This is the meaning of Large Margin classification, from a Functional perspective.



Concept of Large Margin Classification

- Now let us look at this issue from a completely different perspective
- The figure below illustrates data points from 2 different classes (analogous to y = 0 and y = 1) nicely separated by a plane (line in 2-D) serving as a boundary
- Let us compare the pts. A, B and C. Point C is very close to the boundary, and a minor change in inputs of C or in the boundary slope will push it into the other side resulting in misclassification
- Pt. A is very far from the boundary and minor changes will still keep it securely on the correct side of classification. So we can say its reliability and classification confidence is much higher
- This is the meaning of Large Margin Classification, from a Geometrical perspective.



SVM - Terminology

- Now that we are clear on the concept of Large Margin Classification, let us turn to Support Vector Machines
- Terminology (as distinct from those developed for *Supervised Learning*):
  - First, we still deal with two classes, but these are for y = +1, and y = -1, in place of y = 0
  - The hypothesis is expressed in terms of **w** (vector) and b (scalar), instead of θ, and more specifically

    $h_{w,b}(x) = g(w^T x + b)$. You will see later why we have introduced a *b* here, just note that this replaces the θ_0 we used to have earlier
  - Further, the function g(.) is not the logistic function (obvious because sigmoid is limited on the lower side at 0), but not even the tanh function, and has only two values: $g(z) = 1$ if $z \geq 0$, and $g(z) = -1$ if $z < 0$,

    and it does not take any intermediate values.

SVM

Functional Margin of SVM

- For a training sample (x^i, y^i) from a set of *m* such samples, the Functional Margin is defined as

$$\hat{\gamma}^i = y^i \left(w^T x^i + b \right) \quad \ldots \quad (1)$$

- Now y^i can be either + or -1, and ($w^T x^i + b$) if correctly classified will be correspondingly either ≥ 0 or < 0, it follows then that $\hat{\gamma}^i > 0$ always, under correct classification
- We define the *functional margin* of (w, b) for a given data set (x^i, y^i) with i = 1, .., m, as $\hat{\gamma} = \min_{i=1,..,m} \hat{\gamma}^i$ … (2)
- An interesting property of g(.):
  - Since $g(w^T x + b)$ is either +1 when $w^T x + b > 0$, or -1 when $w^T x + b < 0$, it follows that if we replace (w, b) with (2w, 2b), the value of g(.) remains unchanged. This is true for any (nw, nb) for a +(ve) real *n*. But the functional margin $\hat{\gamma}^i$ - refer eq. (1) – amplifies by that factor *n* !
  - *So we can arbitrarily change the functional margin – without affecting the original classification - just by appropriately scaling (w, b).*

SVM

Geometric Margin of SVM

- First, we need to take a little deviation into the equation for a plane: In 3-D, the general form of the equation of a plane is:

$$lx + my + nz + d = 0 \quad \ldots \quad (3)$$

where *l, m and n* are the three direction ratios of the normal to the plane, and the corresponding direction cosines (cos of angles made by the normal line against the 3 axes, giving the 3 axial components of the unit normal) are given by

$$\frac{l}{\left(l^2 + m^2 + n^2\right)^{1/2}}, \frac{m}{\left(l^2 + m^2 + n^2\right)^{1/2}}, \quad etc.$$

('d' is the perpendicular distance from the origin to the plane).

- Generalizing from 3-D to a hyper-plane, we can write the general form as $w^T x + b = 0$ … (4)

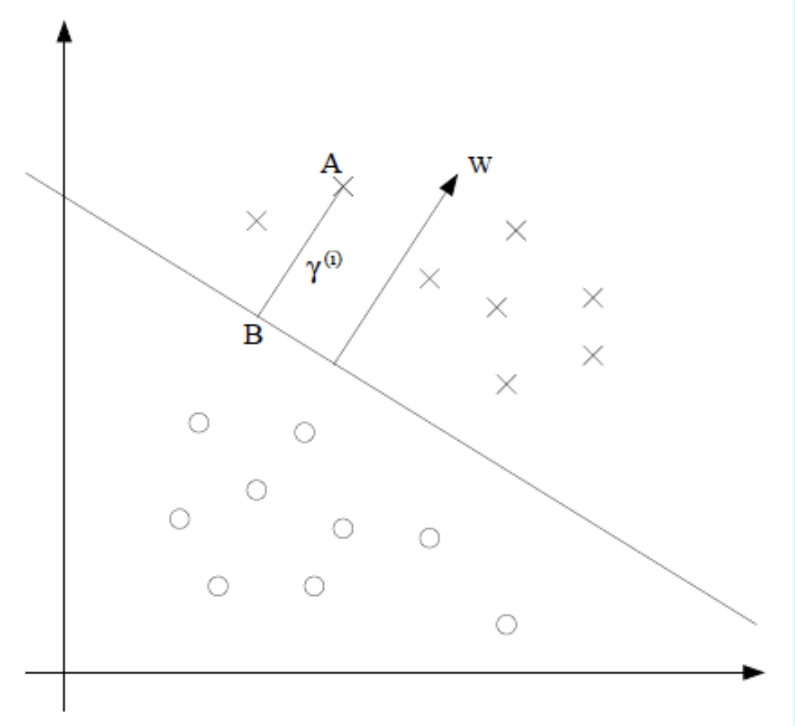
where the direction cosines of the normal are given by the vector

$$\frac{w}{\|w\|} \quad with \ \|w\| = \left(\sum_{i=1}^{n} w_i^2\right)^{1/2}, \text{ n is the spatial dimension} \quad \ldots \quad (5)$$

SVM

Geometric Margin of SVM

- Now let us look at the figure showing two different classes with a separating boundary plane, a pt. A in one class with its perpendicular projection on the boundary plane at B, and a normal vector w to the boundary plane with its unit normal given by w/||w||



- Note that pt. A is actually the vector $\mathbf{x}^i$ where i is simply the identity of that point sample, with $i = 1, .., m$, and the length of vector AB is γ^i, where γ, a scalar, symbolically denotes the geometric margin
- Then we can represent vector B = vector A - γ^i x *unit normal vector*, or $$\mathbf{B} = \mathbf{x}^i - \gamma^i \left(\mathbf{w} / \|\mathbf{w}\| \right) \quad \ldots \quad (6)$$ once we are clear about the vectors, we will get rid of the **bold**.

SVM

Geometric Margin of SVM

- But pt. B lies on the plane hence satisfies eq. (4), so

$$w^T\left(x^i-\gamma^i\left(w/\|w\|\right)\right)+b=0 \quad \ldots \quad (7)$$

- Solving for γ^i yields

$$\gamma^i=\left(\frac{w}{\|w\|}\right)^T x^i+\frac{b}{\|w\|} \quad \ldots \quad (8)$$

- The above example was worked out for the case where $y^i > 0$ (= +1)
- One may analogously work out for the -(ve) example, where RHS of (8) is -(ve), and both can be expressed in an unified fashion as

$$\gamma^i=y^i\left(\left(\frac{w}{\|w\|}\right)^T x^i+\frac{b}{\|w\|}\right) \quad \ldots \quad (9)$$

- Finally, the Geometric Margin of (w, b) with respect to a data set (x^i, y^i) with i = 1, .., m is defined by $\gamma=\min_{i=1,..,m}\gamma^i$ ... (10)

SVM: Reconciliation of Functional and Geometric Margins

We reproduce below both Functional and Geometric Margins

- The Functional Margin of SVM was defined as:

$$\hat{\gamma}^i = y^i \left(w^T x^i + b \right) \quad \ldots \text{ from eq. (1)}$$

- And we saw the Geometric Margin in the last slide:

$$\gamma^i = y^i \left(\left(\frac{w}{\|w\|} \right)^T x^i + \frac{b}{\|w\|} \right) \quad \ldots \text{ eq. (9)}$$

- It would be obvious that:

a) $\hat{\gamma} = \gamma \|w\|$ (note that we have removed the index i) … (11)

b) If $\|w\| = 1$, Functional and Geometric Margins are same

c) The Geometric Margin is invariant to scaling of w or b, unlike the Functional Margin which scaled in proportion to this scaling.

SVM: Objective Function for attaining optimum boundary

- Recall that our original problem in SVM was to obtain *that boundary plane between two classes that provides the Largest Margin Classification*
- The parameters which defined this plane were (w, b), so effectively we need to define the optimization problem in a manner which leads to synthesis of these parameters
- Prima facie the apparent way to define this problem would be :

$$\underset{w,\,b}{Maximize}\ \gamma$$

$$subject\ to\ \ y^i\left(w^T x^i + b\right) \geq \gamma,\ i = 1,..,m \qquad \ldots \quad (12)$$

$$and\ \ \|w\| = 1$$

- Note that this is a maximization problem, and the $\|w\| = 1$ constraint that ensures equivalence of the Geometric and Functional margins complicates the problem.

SVM: Objective Function for attaining optimum boundary

- We eliminate the $\|w\| = 1$ constraint by mildly reformulating the optimization problem as

$$\underset{w,\, b}{Maximize} \quad \frac{\hat{\gamma}}{\|w\|}$$

$$subject\ to \quad y^i \left(w^T x^i + b \right) \geq \hat{\gamma}, \ i = 1, .., m \qquad \ldots \quad (13)$$

- Notice that the eq. (11) relating the two (F and G) margins allowed us to express the objective function in the above form, i.e. in terms of the Functional Margin. But the function is still in a non-convex form.
- Now recall an interesting property of Functional Margins : we can scale (w, b) to correspondingly scale this F-Margin
- Hence, at a particular scale, we can attain the value $\hat{\gamma} = 1$. (For a conceptual analogy, you can think of the constant term in an indefinite integral.)

SVM: Objective Function for attaining optimum boundary

- On making the substitution $\hat{\gamma} = 1$, our optimization problem changes to

$$\underset{w,\, b}{Maximize} \ \frac{1}{\|w\|}$$

$$subject\ to\ \ y^i\left(w^T x^i + b\right) \geq 1,\ i = 1,..,m \qquad \ldots \quad (14)$$

- The maximization problem in (14) can be converted into a quadratic minimization problem

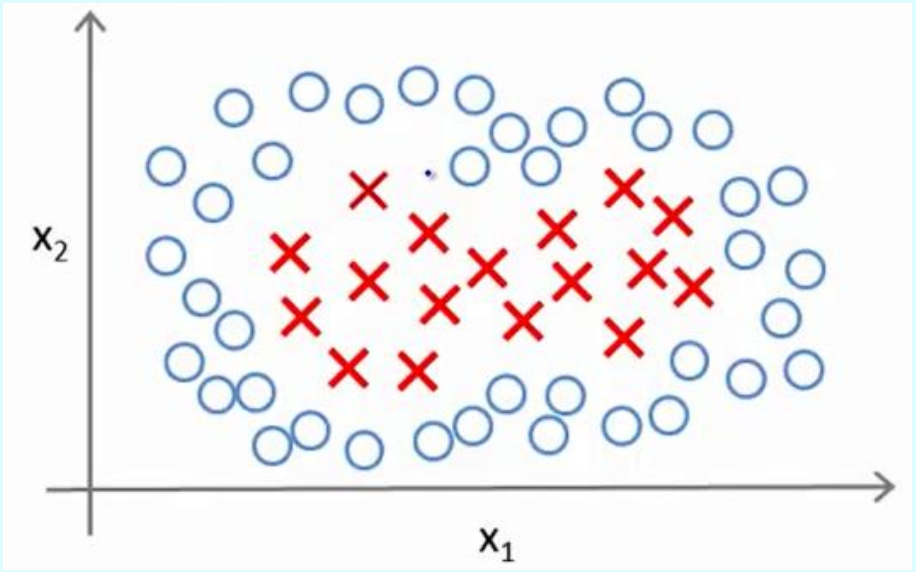
$$\underset{w,\, b}{Minimize} \ \frac{1}{2}\|w\|^2$$

$$subject\ to\ \ y^i\left(w^T x^i + b\right) \geq 1,\ i = 1,..,m \qquad \ldots \quad (15)$$

- The objective function in (15) is a convex quadratic function and can be efficiently solved using Quadratic Programming algorithms
- This brings us to the end of Large Margin Classification of Linearly separable problems using SVM
- Next we will see how to perform Large Margin Classification of non-linearly separable problems with SVMs.

Problems with Non-linear boundaries

- The illustration shows distribution of two classes with a non-linear separation boundary



In the supervised learning framework, we would express a hypothesis $h_\theta(x)$ for a nonlinear problem as

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1 x_2 + \theta_4 x_1^2 + \theta_5 x_2^2 + \ldots\ldots\ldots$$

Further, we would say : Predict y = 1 if $h_\theta(x) \geq 0$, and y = -1 if $h_\theta(x) < 0$.

- Alternately, we could express a hypothesis as

  $h_\theta(f) = \theta_0 + \theta_1 f_1 + \theta_2 f_2 + \theta_3 f_3 + \theta_4 f_4 + \theta_5 f_5 + \ldots$

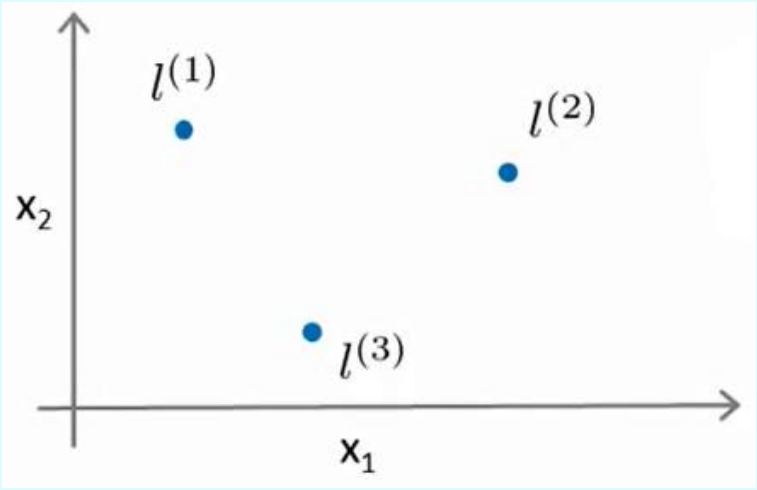
  where $f_1 = x_1$, $f_2 = x_2$, $f_3 = x_1 x_2$, $f_4 = x_1^2$ and so on, with similar class prediction conditions

- Could there be a different/better choice of features f_1, f_2, f_3, etc.?

- Most illustrations in these slides from Andrew Ng's notes

The Kernel Approach

- The concept of *Kernels* is based on the underlying concept of *landmarks*. Let us see with an example



In the illustrated 2-D space, let us choose 3 specific points l^1, l^2, l^3. We will call them landmarks. Why and how we arrive at specific landmarks, we shall discuss later.

- Now let us take any new point of our interest, *x*.
- We want to evaluate the *Similarity* between x, and each of the landmarks l^i. Let us define the similarity in terms of the function:

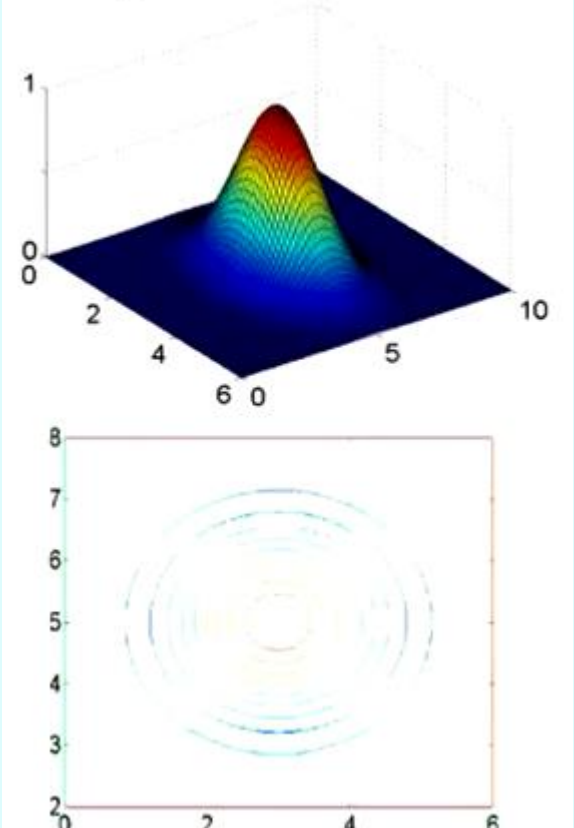
$$Similarity(x, l^i) = e^{\frac{-\|x-l^i\|^2}{2\sigma^2}} \equiv \exp\left(-\frac{\|x-l^i\|^2}{2\sigma^2}\right) \quad \ldots \quad (16)$$

- The Kernel function is nothing but this Similarity function, and the specific Kernel function above is called the *Gaussian Kernel*.

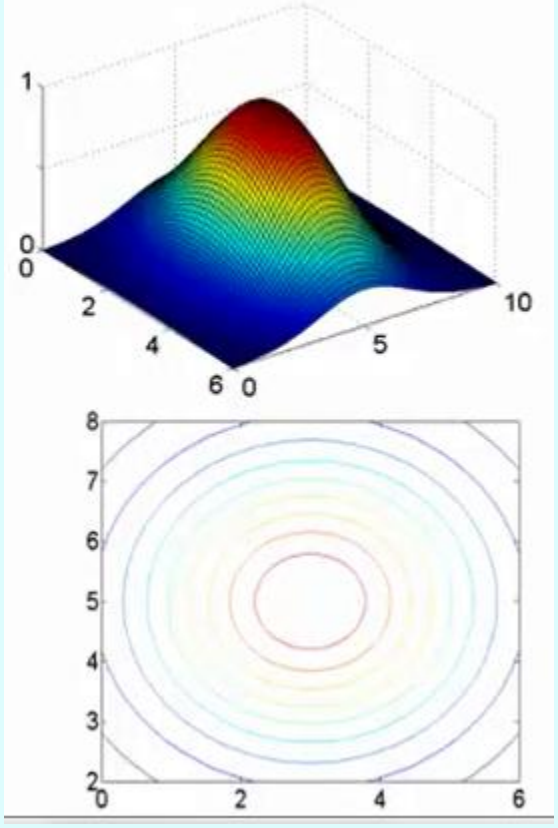
Investigating the behaviour of the Gaussian Kernel

- Let us investigate the behaviour of the Gaussian Kernel for a specific landmark point with respect to variations in the field point x, and the effect of σ. Let this landmark point be $l^1 = [3, 5]^T$, and recall the Kernel function $\exp\left(-\left\|x - l^i\right\|^2 / 2\sigma^2\right)$
- The function behaviour for different values of σ are shown below:

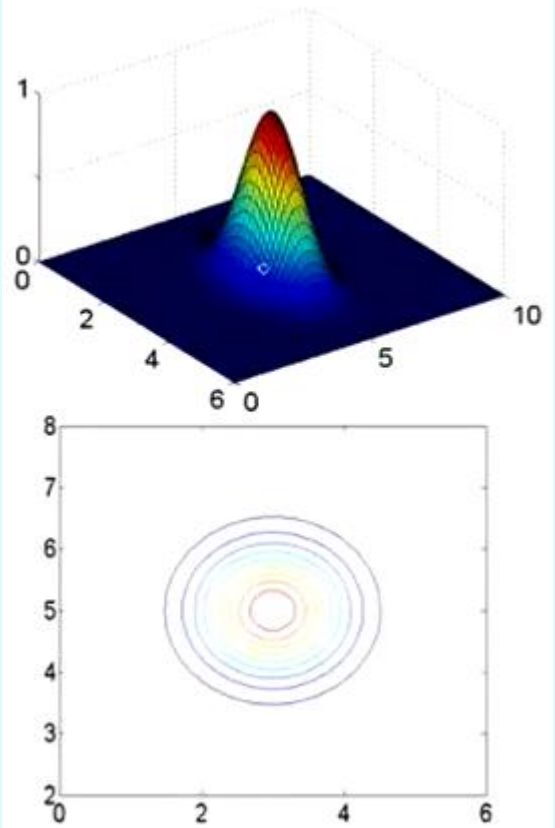
$\sigma^2 = 1$



$\sigma^2 = 3$

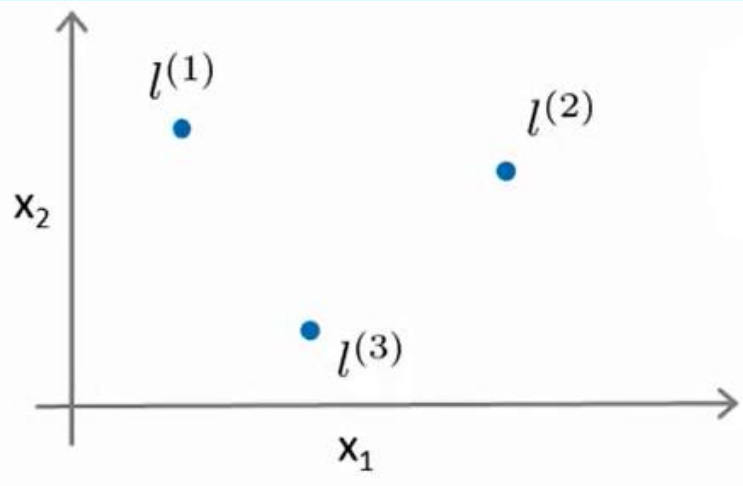


$\sigma^2 = 0.5$

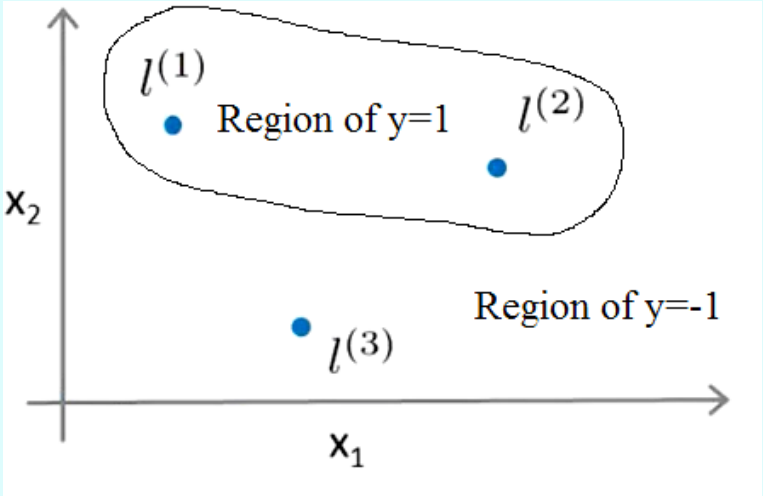


Generating Features using Landmarks

- Recall that we could express a hypothesis for a nonlinear boundary classification problem as $h_\theta(f) = \theta_0 + \theta_1 f_1 + \theta_2 f_2 + \theta_3 f_3 + \theta_4 f_4 + \theta_5 f_5 + \ldots$ with $f_1 = x_1$, $f_2 = x_2$, $f_3 = x_1 x_2$, $f_4 = x_1^2$ and so on, with class prediction conditions $y = 1$ for $h_\theta(f) \geq 0$, and $y = -1$ for $h_\theta(f) < 0$
- Could there be an alternate / better choice of features f_1, f_2, f_3, etc.?
- *Let us choose each f_i to be the Kernel function around landmark l^i, i.e. f_i = Similarity(x, l^i)* $\equiv \exp\left(-\frac{\|x - l^i\|^2}{2\sigma^2}\right)$
- In that case, with the 3 landmarks, $h_\theta(f)$ becomes $\theta_0 + \theta_1 f_1 + \theta_2 f_2 + \theta_3 f_3$
- Suppose we take $\theta_0 = -0.5$, $\theta_1 = 1$, $\theta_2 = 1$ and $\theta_3 = 0$

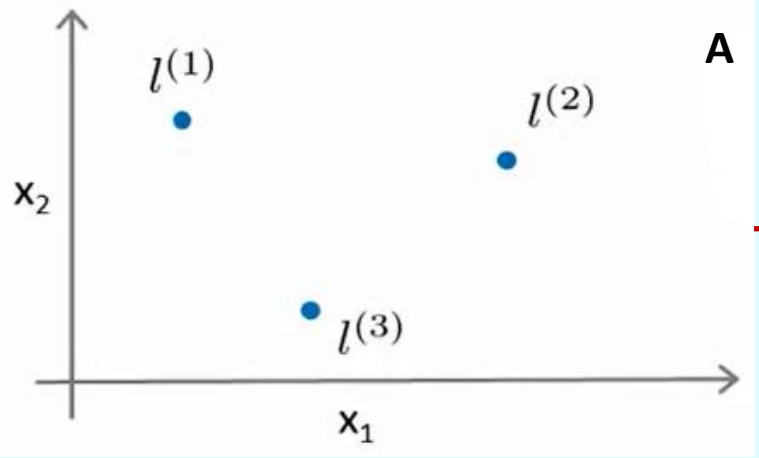


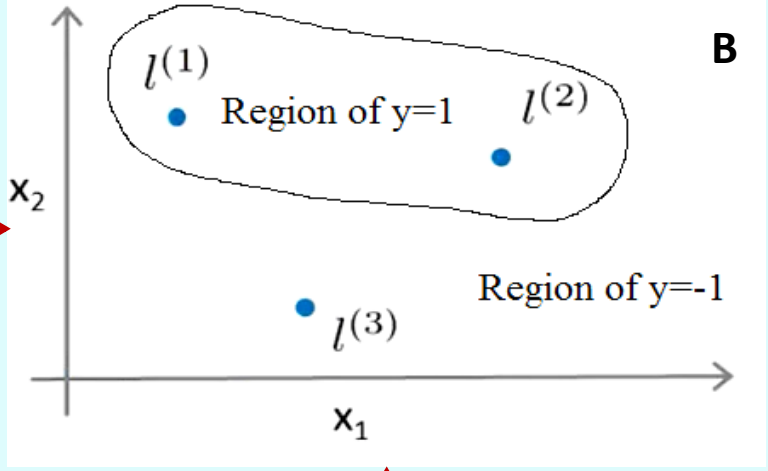
Then using detailed observations we can find that the class y=1 encompasses the region around pts. l^1 and l^2, while the rest is in class y=-1

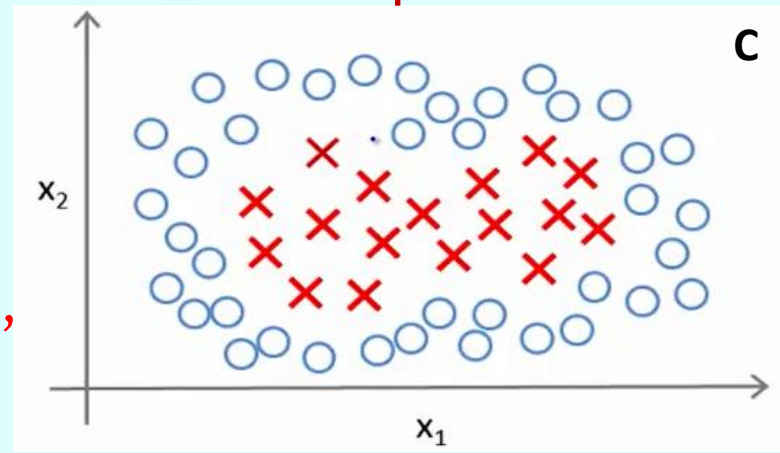


Expressing the Hypothesis using Landmarks

- Hence, using the Gaussian Kernel Function, and the 3 selected points (Fig. A) as Landmarks, we have been able to extract the nonlinear boundary function (Fig. B), that classifies the two types of points shown in Fig. C.







- $h_\theta(f) = \theta_0 + \theta_1 f_1 + \theta_2 f_2 + \theta_3 f_3$ with
- $\theta_0 = -0.5$, $\theta_1 = 1$, $\theta_2 = 1$ and $\theta_3 = 0$
- Now the Million Dollar Question:
- How to arrive at a choice of landmark pts. like, e.g., l^1, l^2, l^3, and a corr. choice of θ's, for accurately classifying some gn. dstbn?

SVM with Kernels for non-linear classification

- Given a training data set (x^i, y^i), i = 1, .., m
- We choose $l^1 = x^1$, $l^2 = x^2$, …., $l^m = x^m$!!
- It then follows that for any new example *x* (that we may want to classify), feature f_1 = similarity(x, l^1), ….., f_m = similarity (x, l^m)
- In particular, for any training example (x^i, y^i), we have

$f_1^i = sim(x^i, l^1)$

$f_2^i = sim(x^i, l^2)$

.

.

$f_i^i = sim(x^i, l^i) = 1$

.

$f_m^i = sim(x^i, l^m)$

⇓

f^i = sim(x^i, l), where **f^i, l** are vectors with m components.

And the final optimization problem changes to

$$\underset{w,\,b}{Minimize}\ \frac{1}{2}\|w\|^2$$

$$subject\ to\ \ y^i\left(w^T f^i + b\right) \geq 1,\ i = 1,..,m \ \ldots\ (17)$$

One point on regularization. Notice that the normalization coefficient σ behaves as an effective regularizer - small σ implies high variance and low bias – and higher the σ, more the regularization effect.

Note: Role of w′s here are same as θ′s in previous slide.

THANK YOU