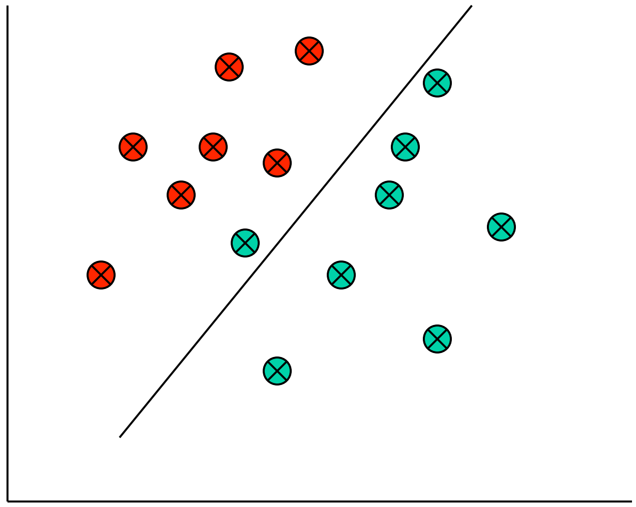
Basis Expansion and Nonlinear SVM

Kai Yu

Linear Classifiers

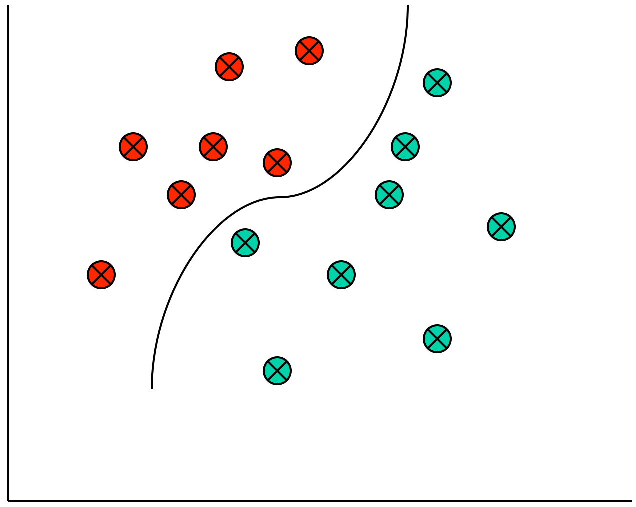


$$f(x) = w^\top x + b$$

$$z(x) = \mathrm{sign}(f(x))$$

- Help to learn more general cases, e.g., nonlinear models

Nonlinear Classifiers via Basis Expansion



$$f(x) = w^\top h(x) + b$$

$$z(x) = \text{sign}(f(x))$$

- Nonlinear basis functions $h(x)=[h_1(x), h_2(x), \ldots, h_m(x)]$
- $f(x) = w^T x+b$ is a special case where $h(x)=x$
- This explains a lot of classification models, including SVMs.

Outline

- Representation theorem
- Kernel trick
- Understand regularization
- Nonlinear logistic regression
- General basis expansion functions
- Summary

Review the QP for linear SVMs

- After a lot of "stuff", we obtain the Lagrange dual

$$L_D = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{i'=1}^{N} \alpha_i \alpha_{i'} y_i y_{i'} x_i^T x_{i'}$$

- The solution has the form

$$w = \sum_{i=1}^{N} \alpha_i y_i x_i$$

- In other words, the solution w is in

$$\mathrm{span}(x_1, x_2, \ldots, x_N)$$

A more general result – RKHS representation theorem (Wahba, 1971)

- In its simplest form, $L(w^T x, y)$ is covex w.r.t. w, the solution of

$$\min_w \sum_{i=1}^{N} L(w^T x_i, y_i) + \lambda \|w\|^2$$

has the form
$$w = \sum_{i=1}^{N} \alpha_i x_i$$

- Proof sketch …
- Note: the conclusion is general, not only for SVMs.

For general basis expansion functions

The solution of

$$\min_w \sum_{i=1}^{N} L(w^\top h(x_i), y_i) + \lambda \|w\|^2$$

has the form

$$w = \sum_{i=1}^{N} \alpha_i h(x_i)$$

Outline

- Representation theorem
- Kernel trick
- Understand regularization
- Nonlinear logistic regression
- General basis expansion functions
- Summary

Kernel

- Define the Mercer kernel as

$$k(x_i, x_j) = h(x_i)^\top h(x_j)$$

Kernel trick

- Apply the representation theorem

$$w = \sum_{i=1}^{N} \alpha_i h(x_i)$$

- we have

$$f(x) = \sum_{i=1}^{N} \alpha_i k(x_i, x) \qquad \|w\|^2 = \sum_{i,j=1}^{N} \alpha_i \alpha_j k(x_i, x_j) = \alpha^T K \alpha$$

$$\min_{\alpha} \sum_{i=1}^{N} L\left(\sum_{i=1}^{N} \alpha_i k(x_i, x), y_i\right) + \lambda \alpha^\top K \alpha$$

Primal and Kernel formulations

$$\min_{w} \sum_{i=1}^{N} L\left(w^\top h(x), y_i\right) + \lambda \|w\|^2$$

$$k(x_i, x_j) = h(x_i)^\top h(x_j)$$

$$\min_{\alpha} \sum_{i=1}^{N} L\left(\sum_{i=1}^{N} \alpha_i k(x_i, x), y_i\right) + \lambda \alpha^\top K \alpha$$

- Given a kernel, we don't even need h(x)! …really?

Popular kernels

- k(x,x') is a symmetric, positive (semi-) definite function

$$d\text{th deg. poly.: } K(x,x') = (1 + \langle x, x' \rangle)^d$$

$$\text{radial basis: } K(x,x') = \exp(-\|x - x'\|^2/c)$$

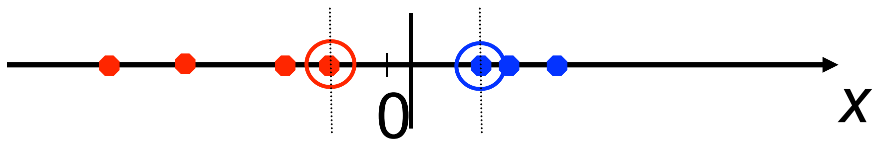
- Example:

$$\begin{aligned} K(x,x') &= (1 + \langle x, x' \rangle)^2 \\ &= (1 + x_1 x_1' + x_2 x_2')^2 \\ &= 1 + 2x_1 x_1' + 2x_2 x_2' + (x_1 x_1')^2 + (x_2 x_2')^2 + 2x_1 x_1' x_2 x_2' \end{aligned}$$

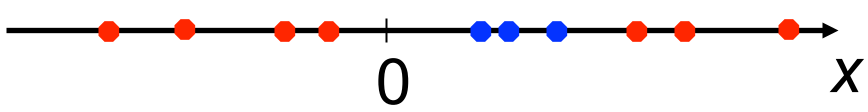
$h_1(x) = 1,\ h_2(x) = \sqrt{2}x_1,\ h_3(x) = \sqrt{2}x_2,\ h_4(x) = x_1^2,\ h_5(x) = x_2^2,$
and $h_6(x) = \sqrt{2}x_1 x_2,$

Non-linear feature mapping

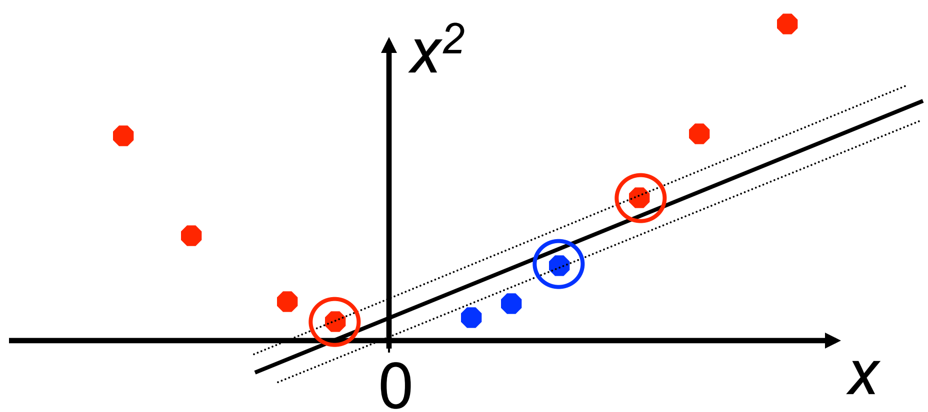
- Datasets that are linearly separable



- But what if the dataset is just too hard?

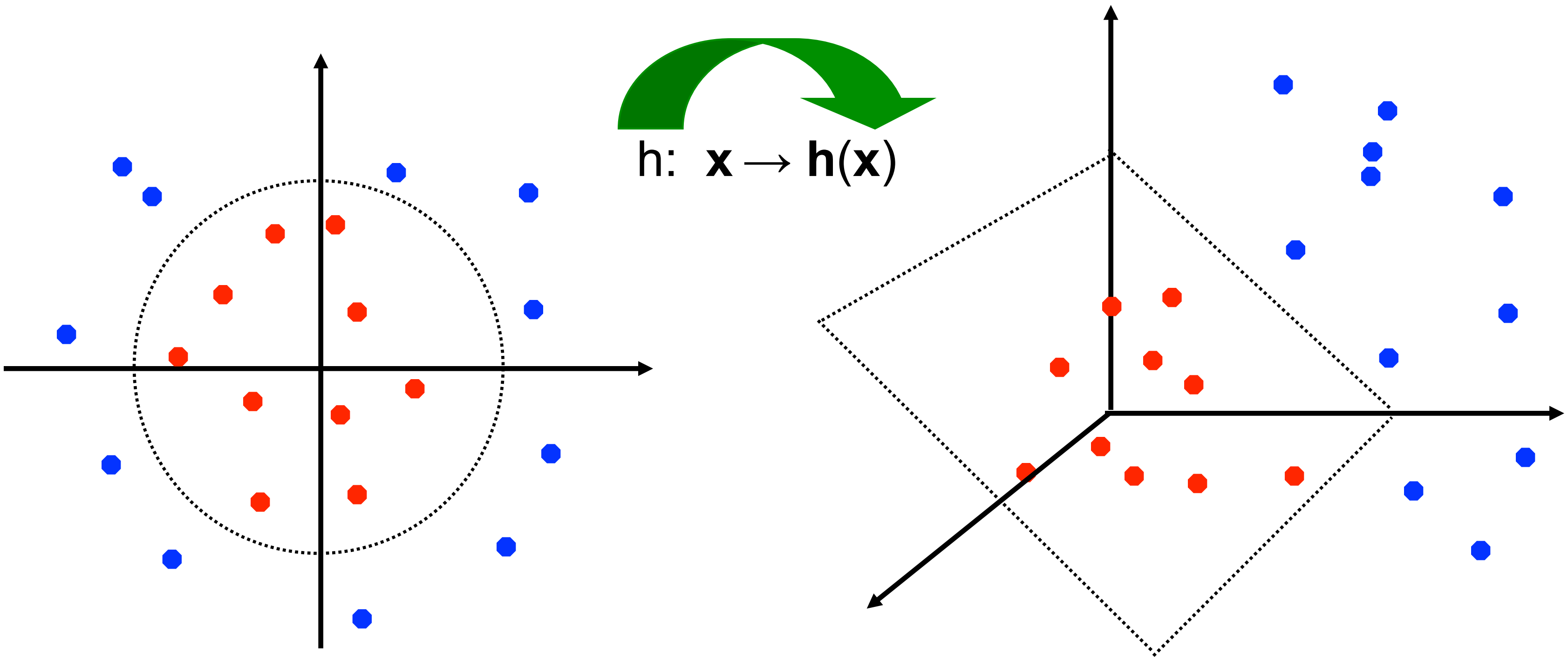


- How about mapping data to a higher-dimensional space:



Nonlinear feature mapping

- General idea: the original feature space can always be mapped to some higher-dimensional feature space where the training set is separable:



Outline

- Representation theorem
- Kernel trick
- Understand regularization
- Nonlinear logistic regression
- General basis expansion functions
- Summary

Various equivalent formulations

- Parametric form

$$\min_{w} \sum_{i=1}^{N} L\left(w^\top h(x), y_i\right) + \lambda \|w\|^2$$

- Dual form

$$\min_{\alpha} \sum_{i=1}^{N} L\left(\sum_{i=1}^{N} \alpha_i k(x_i, x), y_i\right) + \lambda \alpha^\top K \alpha$$

- Nonparametric form

$$\min_{f} \sum_{i=1}^{N} L(f(x_i), y_i) + \lambda \|f\|^2_{\mathcal{H}_k}$$

Various equivalent formulations

- Parametric form

$$\min_{w} \sum_{i=1}^{N} L\left(w^\top h(x), y_i\right) + \lambda \|w\|^2$$

- Dual form

$$\min_{\alpha} \sum_{i=1}^{N} L\left(\sum_{i=1}^{N} \alpha_i k(x_i, x), y_i\right) + \lambda \alpha^\top K \alpha$$

- Nonparametric form

Telling what kind of f(x) is preferred

$$\min_{f} \sum_{i=1}^{N} L(f(x_i), y_i) + \lambda \|f\|_{\mathcal{H}_k}^2$$

Regularization induced by kernel (or basis functions)

$$\text{Eigen expansion: } K(x,y) = \sum_{i=1}^{\infty} \gamma_i \phi_i(x)\phi_i(y)$$

$$f(x) = \sum_{i=1}^{\infty} c_i \phi_i(x)$$

- Desired kernel is a smoothing operator, smoother eigenfunctions ϕ_i tend to have larger eigenvalues γ_i

$$||f||^2_{\mathcal{H}_K} \stackrel{\text{def}}{=} \sum_{i=1}^{\infty} c_i^2/\gamma_i$$

- What does this mean ?

Understand regularization

- If push down this regularization term

$$||f||^2_{\mathcal{H}_K} \stackrel{\text{def}}{=} \sum_{i=1}^{\infty} c_i^2 / \gamma_i$$

- In f(x), minor components $\phi_i(x)$ with smaller γ_i are penalized more heavily. → principle components are preferred in f(x)!

- A desired kernel is a smoothing operator, i.e., principle components are smoother functions → the regularization encourages f(x) to be smooth!

Understanding regularization

$$||f||^2_{\mathcal{H}_K} \overset{\text{def}}{=} \sum_{i=1}^{\infty} c_i^2/\gamma_i$$

- Using what kernel?
- Using what feature (for linear model) ?
- Using what h(x)?
- Using what functional norm $||f||^2_{\mathcal{H}_k}$

All pointing to one thing –
what kind of functions are preferred *apriori*

Outline

- Representation theorem
- Kernel trick
- Understand regularization
- Nonlinear logistic regression
- General basis expansion functions
- Summary

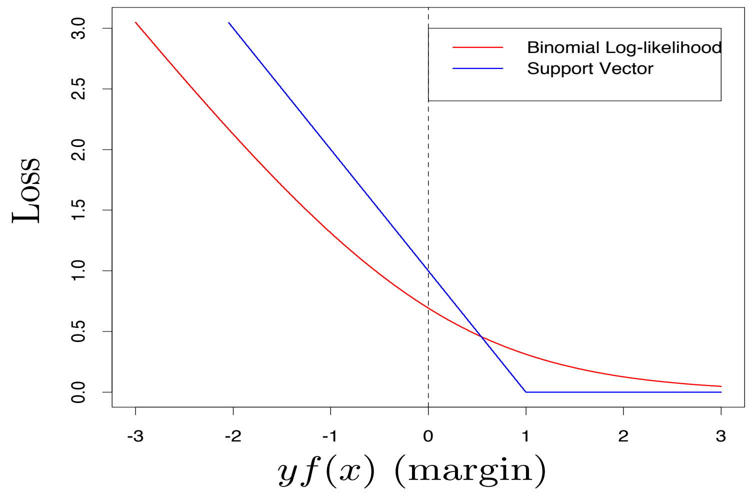
Nonlinear Logistic Regression

So far, things we discussed, including

- representation theorem,
- kernel trick,
- regularization,

are not limited to SVMs. They are all applicable to logistic regression. The only difference is the loss function.

Nonlinear Logistic Regression



- Parametric form

$$\min_f \sum_{i=1}^{N} \ln\left(1 + e^{-y_i w^\top h(x_i)}\right) + \lambda \|w\|^2$$

- Nonparametric form

$$\min_f \sum_{i=1}^{N} \ln\left(1 + e^{-y_i f(x_i)}\right) + \lambda \|f\|_{\mathcal{H}_k}^2$$

Logistic Regression vs. SVM

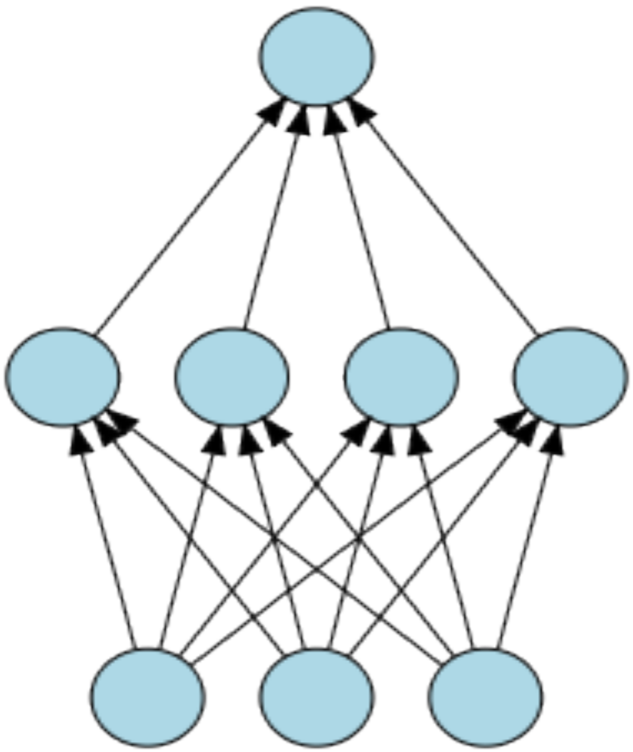
- Both can be linear or nonlinear, parametric or nonparametric, the main difference is the loss;
- They are very similar in performance;
- Outputs probabilities, useful for scoring confidence;
- Logistic regression is easier for multiple classes.
- 10 years ago, one was old, the other is new. Now, both are old.

Outline

- Representation theorem
- Kernel trick
- Understand regularization
- Nonlinear logistic regression
- General basis expansion functions
- Summary

Many known classification models follow a similar structure

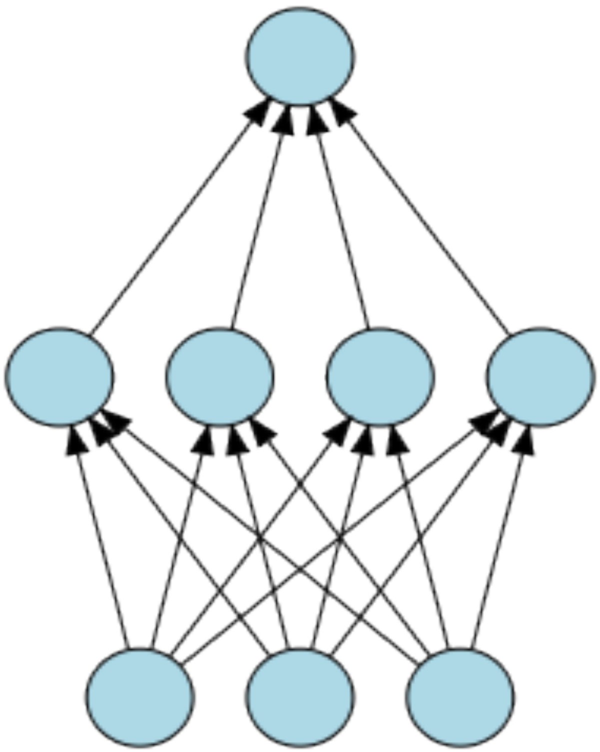
- Neural networks
- RBF networks
- Learning VQ (LVQ)
- Boosting



These models all learn w and h(x) together …

Many known classification models follow a similar structure

- Neural networks
- RBF networks
- Learning VQ (LVQ)
- Boosting
- SVMs
- Linear Classifier
- Logistic Regression
- …



Develop your own stuff !

By deciding

- Which loss function? – hinge, least square, …
- What form of h(x)? – RBF, logistic, tree, …
- Infinite h(x) or h(x)?
- Learning h(x) or not?
- How to optimize? – QP, LBFGS, functional gradient, …

you can obtain various classification algorithms.

Parametric vs. nonparametric models

- h(x) is finite dim, parametric model $f(x)=w^T h(x)$. Training complexity is $O(Nm^3)$

- h(x) is nonlinear and infinite dim, then has to use kernel trick. This is a nonparametric model. The training complexity is around $O(N^3)$

- Nonparametric models, including kernel SVMs, Gaussian processes, Dirichlet processes etc., are elegant in math, but nontrivial for large-scale computation.

Outline

- Representation theorem
- Kernel trick
- Understand regularization
- Nonlinear logistic regression
- General basis expansion functions
- Summary

Summary

- Representation theorem and kernels
- Regularization prefers principle eigenfunctions of the kernel (induced by basis functions)
- Basis expansion - a general framework for classification models, e.g., nonlinear logistic regression, SVMs, …