

Putting the SVM in context

Steve Kroon and Sarel Steel

November 2003

Contents

1. The need for regularization
2. Regularization
3. A formulation for SVM classification
4. SVM as a regularization problem
5. Regularization networks
6. Techniques yielding kernel expansions
7. Other regularization networks
8. Wrapping up
9. Sources

1 Ill-posed problems

- In traditional linear regression, we use the inverse of the matrix $X^T X$
- Sometimes $X^T X$ is *ill-conditioned*, and almost singular, leading to high variance of our estimators
- Particularly common in high-dimensional spaces
- To counteract this, ridge regression is popular - adds a ridge penalty to $X^T X$, reduces the condition number
- Analog for general problems is ill-posedness, particularly common in high-dimension problems

2 Regularization

- We wish to find the $f \in \mathcal{F}$ minimizing

$$\sum_{i=1}^N L(y_i, f(x_i)) + \lambda J(f)$$

- Loss function L , regularizer J , model space \mathcal{F}
- L measures lack of fit of function to data
- J measures complexity of function
- Function chosen from \mathcal{F}
- λ controls tradeoff of these two factors
- Optimal value of λ depends on sample size
- λ analogous to bandwidth in kernel density estimation

- For a given value of λ , there are corresponding values J_λ and L_λ , so that the following two problems are equivalent to the regularization problem:
- Minimize $\sum_{i=1}^N L(y_i, f(x_i))$ subject to $J(f) \leq J_\lambda$; and
- Minimize $J(f)$ subject to $\sum_{i=1}^N L(y_i, f(x_i)) \leq L_\lambda$
- These problems are called the *duals* of each other
- We thus see that λ is effectively a Lagrange multiplier
- It is through these formulations we sometimes discover regularization problems

3 A formulation for SVM classification

- After some modifications, the statement of the SVM classification problem leads to

$$\text{minimize } \frac{1}{2} \|w\|^2 \text{ subject to } \sum_{i=1}^N (1 - y_i f(x_i))_+ \leq L_\lambda$$

- The previous slide tells us that this is then a regularization problem of the form

$$\text{minimize } \sum_{i=1}^N (1 - y_i f(x_i))_+ + \lambda \|w\|^2$$

4 SVM as a regularization problem

- Starting in a linear situation, the origin of SVMs is the perceptron:

$$\text{minimize } \sum_{i=1}^N (-y_i f(x_i))_+$$

- We want our loss to be an upper bound on misclassification error, so modify the loss to the hinge loss. to obtain:

$$\text{minimize } \sum_{i=1}^N (1 - y_i f(x_i))_+$$

- When applying the kernel trick to work in high-dimensional spaces, our problem becomes ill-posed so we introduce a flatness regularizer:

$$\text{minimize } \sum_{i=1}^N (1 - y_i f(x_i))_+ + \lambda \|w\|^2$$

- SVM regression: only slightly different loss function (ε -insensitive loss)

5 Regularization Networks

- For a certain (common) class of regularizers, the optimal solution to the problem can be written as $\sum_{i=1}^N \alpha_i G(x, x_i)$
- This is due to the *Representer Theorem* (Kimeldorf and Wahba, 1971)
- The form of G is solely dependent on the regularizer, the α_i 's depend on λ and the loss function
- For regularizer $\|w\|^2$ of SVM, G turns out to be the kernel
- Hence similar form for kernel FDA and kernel logistic regression (SVM sparsity derived from loss function)

6 Techniques yielding kernel expansions

Technique	Loss function	Formula
SVM Classifier	Hinge loss	$(1 - yf(x))_+$
SVM Regression	ε -insensitive loss	$(y - f(x) - \varepsilon)_+$
Kernel logistic regression	Deviance	$\log(1 + e^{-2yf(x)})$
Kernel Boosting	Exponential	$e^{-yf(x)}$
Kernel FDA Least squares SVM Kernel Ridge regression	Least Squares	$(y - f(x))^2$

7 Other regularization networks

- Radial basis function networks: *e.g.* Gaussian RBF networks use a regularizer $\int e^{-\frac{\|s\|^2}{\beta}} |\tilde{f}(s)|^2 ds$ yielding a *Green's function*
$$G(x, x_i) = e^{-\frac{\|x - x_i\|^2}{2\sigma^2}}$$
- Smoothing splines can use a regularizer such as $\int [f''(t)]^2 dt$ yielding a natural cubic smoothing spline (here
$$G(x, x_i) = (x - x_i)_+^3$$
)
- Regularization networks allow us to search for the optimal solution when it would have been infeasible otherwise

8 Wrapping up

- Regularization not limited to regularization networks
- Besides $\|w\|^2$, $\|w\|_1$ (lasso techniques) is a popular choice of regularizer (impractical for kernel methods, however) for shrinkage and variable selection
- Regularizers emphasising smoothness will typically involve derivatives or Fourier transforms
- Many techniques effectively implement regularization in an *ad hoc* fashion, although not formulated as such, *e.g.* neural networks (weight decay) and best subset selection techniques for variable selection

- Potential use for:
 - studying loss functions and regularizer independently
 - studying interaction of these components
 - new techniques by combining existing components
 - can obtain results quickly for new techniques by showing they belong to the framework
- Example: the regularizer of SVMs allows its kernel expansion, while its loss function encourages sparsity. Kernel logistic regression maintains the kernel expansion, loses the sparsity and gains the ability to predict probabilities from its loss function.

9 Sources

- Many thanks to Sarel Steel for fruitful discussions
- Charles Chui - An Introduction to Wavelets
- Gerald Folland - Fourier Analysis and its Applications
- Federico Girosi, Michael Jones and Tomaso Poggio - Regularization Theory and Neural Networks Architecture (Neural Computation, vol. 7, pp. 219–269)
- Trevor Hastie, Robert Tibshirani and Jerome Friedman - The Elements of Statistical Learning
- Simon Haykin - Neural Networks
- Sebastian Mika - Kernel Fisher Discriminants (Ph. D. thesis, Technical University of Berlin)

- Alex Smola - A Tutorial on Support Vector Regression
(NeuroCOLT2 Technical Report NC2-TR-1998-030)
- Alex Smola and Bernhard Schölkopf - Learning with Kernels
- Vladimir Vapnik - The Nature of Statistical Learning Theory
- Grace Wahba - Spline Models for Observational Data
- Andrew Webb - Statistical Pattern Recognition