Lecture 7: Tuning hyperparameters using cross validation

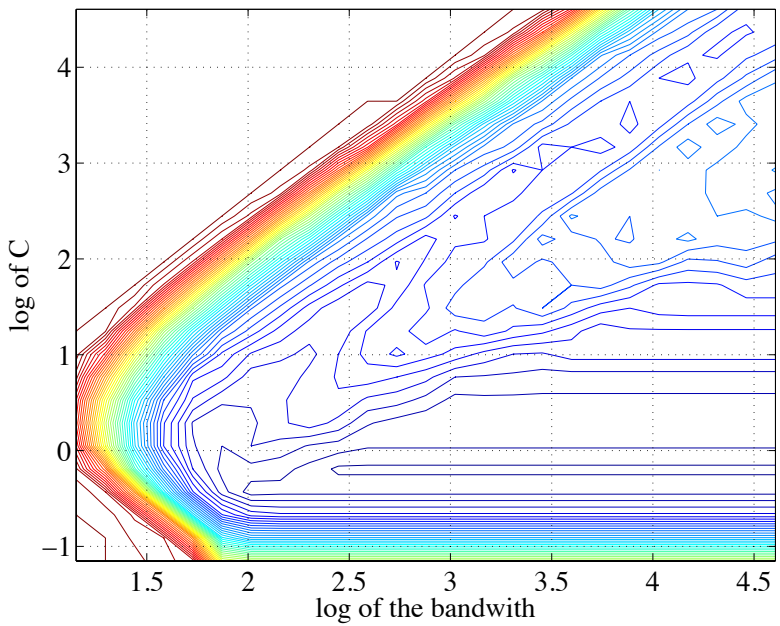
Stéphane Canu
stephane.canu@litislab.eu

Sao Paulo 2014

April 4, 2014

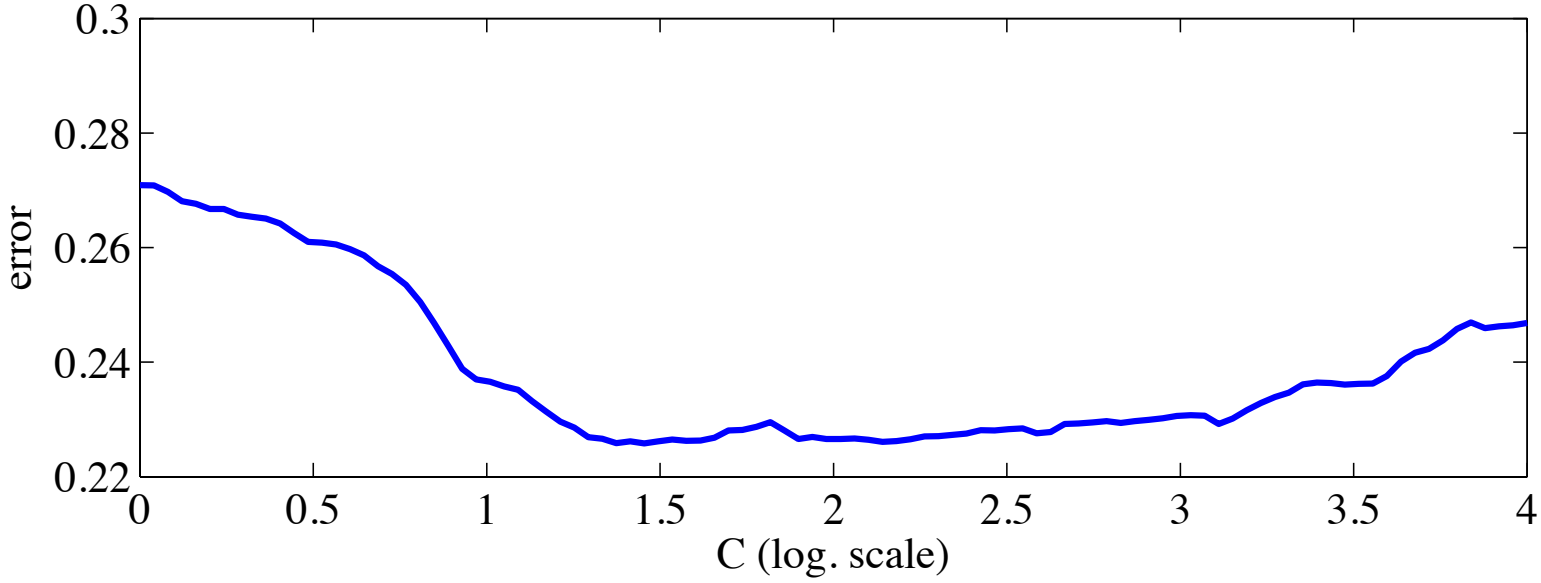
Roadmap

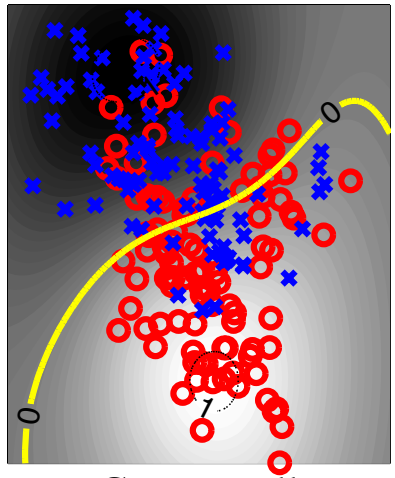
1. Tuning hyperparameters
   - Motivation
   - Machine learning without data
   - Assessing the quality of a trained SVM
   - Model selection



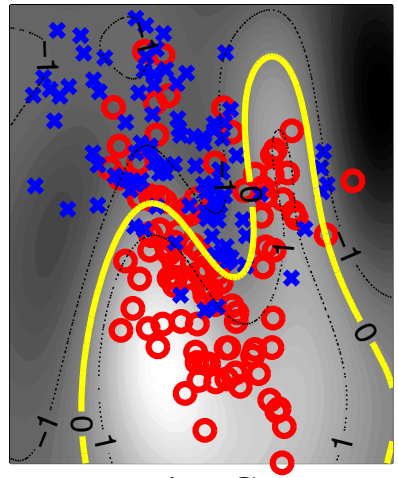
"Evaluation is the key to making real progress in data mining", [Witten & Frank, 2005], p.143 (from N. Japkowicz & M. Shah ICML 2012 tutorial)

Motivation: the influence of C on SVM

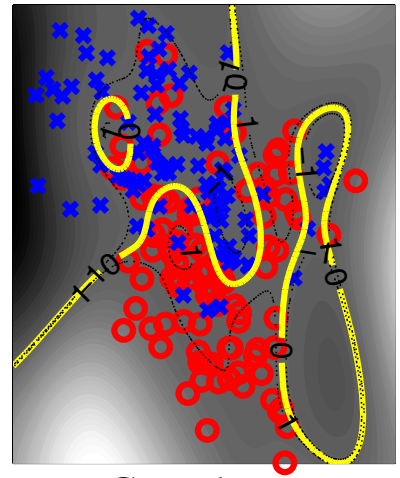




C too small



nice C



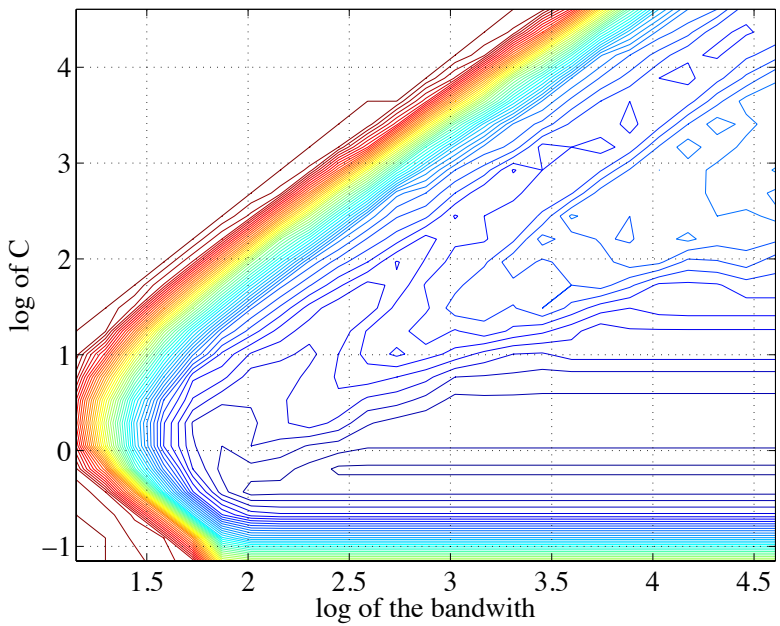
C too large

Motivation:

Need for model selection (tuning the hyper parameters)
Require a good estimation of the performance on future data Choose a relevant performance measure

Roadmap

1. Tuning hyperparameters
   - Motivation
   - Machine learning without data
   - Assessing the quality of a trained SVM
   - Model selection



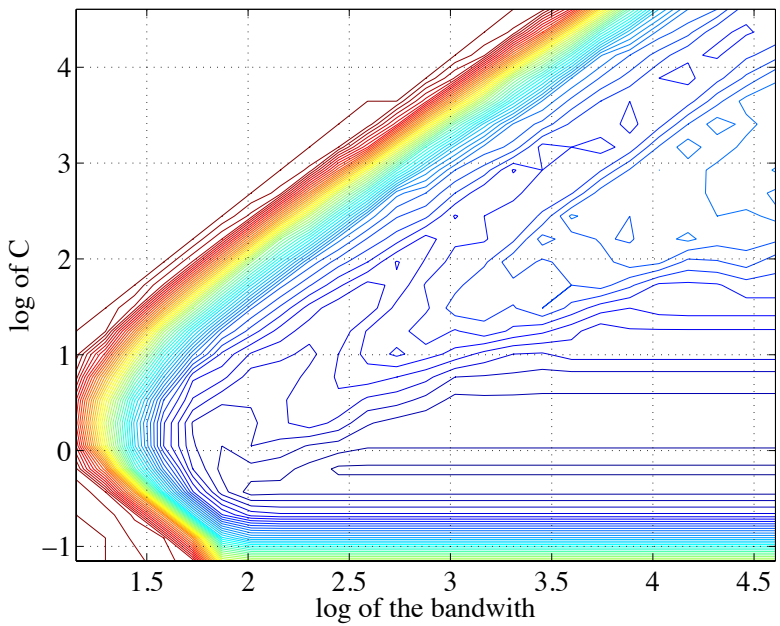
"Evaluation is the key to making real progress in data mining", [Witten & Frank, 2005], p.143 (from N. Japkowicz & M. Shah ICML 2012 tutorial)

Machine learning without data

minimizing $\mathbb{P}(\textit{error})$

Roadmap

1. Tuning hyperparameters
   - Motivation
   - Machine learning without data
   - Assessing the quality of a trained SVM
   - Model selection

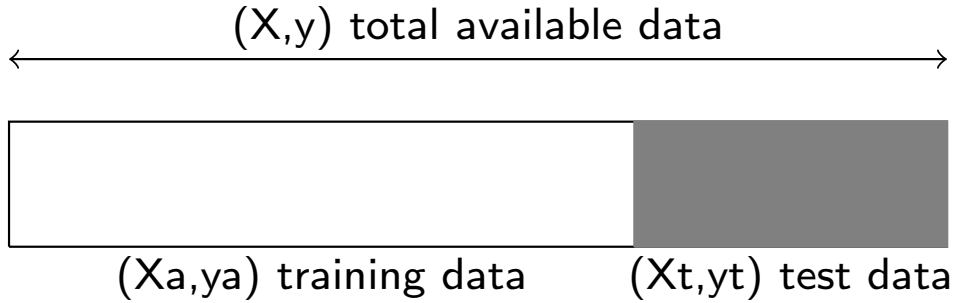


"Evaluation is the key to making real progress in data mining", [Witten & Frank, 2005], p.143 (from N. Japkowicz & M. Shah ICML 2012 tutorial)

Training and test data

Split dataset into two groups randomly picked (hold out strategy)

- Training set: used to train the classifier
- Test set: used to estimate the error rate of the trained classifier



$(Xa, ya, Xt, yt) \leftarrow split(X, y, option = \frac{1}{3})$

- Generally, the larger the training data the better the classifier
- The larger the test data the more accurate the error estimate

Assessing the quality of a trained SVM: minimum error rate

Definition (The confusion matrix)

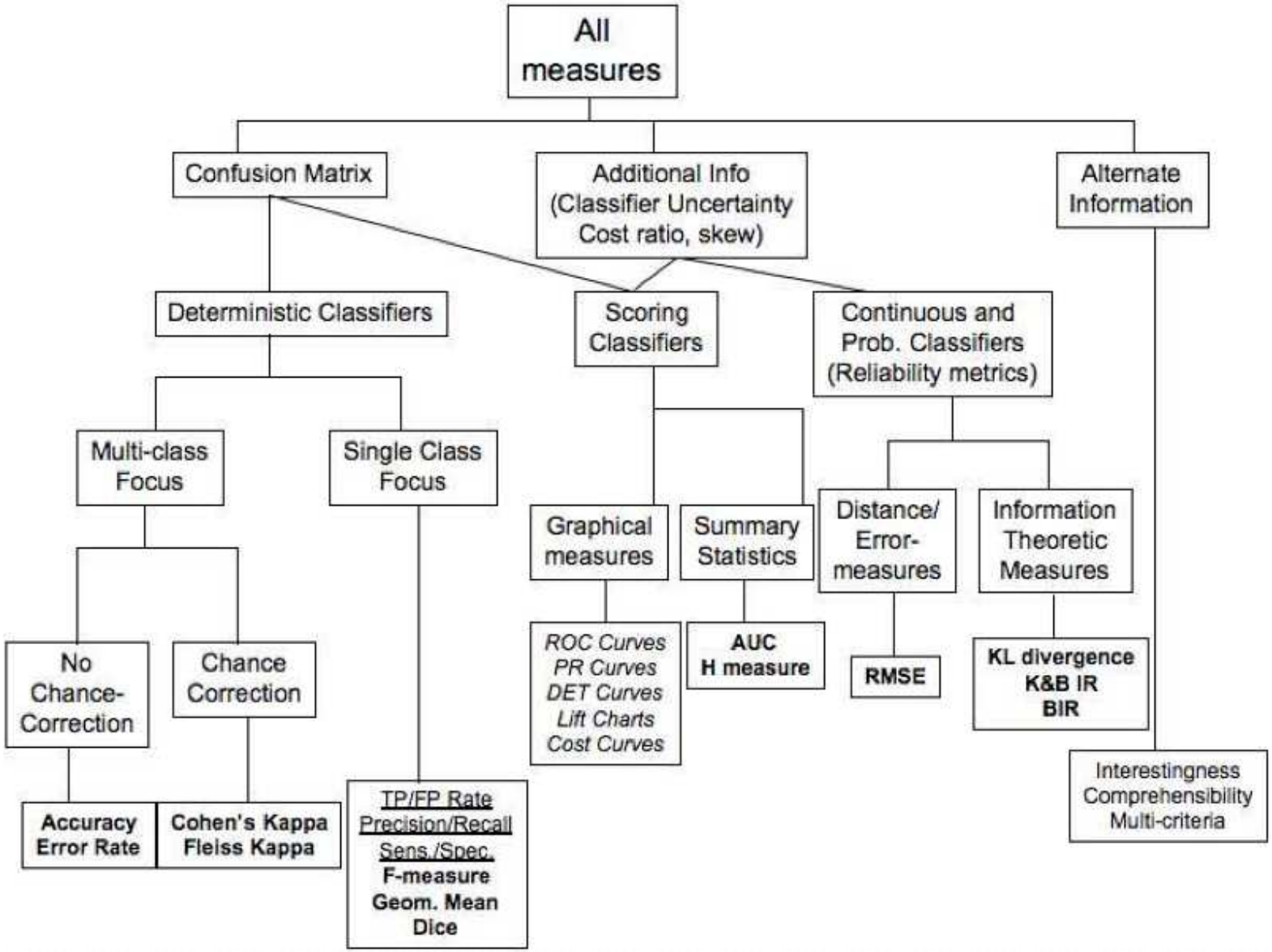
A matrix showing the predicted and actual classifications. A confusion matrix is of size $L \times L$, where L is the number of different classes.

Observed / predicted	Positive	Negative
positive	a	b
negative	c	d

$$\text{Error rate} = 1 - \text{Accuracy} = \frac{b+c}{a+b+c+d} = \frac{b+c}{n} = 1 - \frac{a+d}{n}$$

True positive rate (Recall, Sensitivity) d/(c+d).
True negative rate (Specificity) a/(a+b).
Precision, False positive rate, False negative rate...

Other performances measures



N. Japkowicz & M. Shah, "Evaluating Learning Algorithms: A Classification Perspective", Cambridge University Press, 2011

The learning equation

$$\text{Learning} = \text{training} + \text{testing} + \text{tuning}$$

Table: my experimental error rates

	State of the art	my new method	Bayes error
problem 1	10% $\pm$ 1.25	8.5% $\pm$ .5	
problem 2	5 % (.25)	4 % (.5)	

is my new method good for problem 1?

The learning equation

$$\text{Learning} = \text{training} + \text{testing} + \text{tuning}$$

Table: my experimental error rates

	State of the art	my new method	Bayes error
problem 1	10% $\pm$ 1.25	8.5% $\pm$ .5	11 %
problem 2	5 % (.25)	4 % (.5)	2 %

is my new method good for problem 1?

Error bars on Bernouilli trials

$$\text{Error rate} = \widehat{p} \qquad \mathcal{B}(p)$$

with confidence α: (Normal approximation interval)

$$p = \mathbb{P}(\mathit{error}) \text{ in } \widehat{p} \pm u_{1-\alpha/2}\sqrt{\frac{\widehat{p}\,(1-\widehat{p})}{n_t}}$$

with confidence α: (improved approximation)

$$p = \mathbb{P}(\mathit{error}) \text{ in } \frac{1}{1+\frac{1}{K}u_{1-\alpha/2}^2}\left(\widehat{p} \pm u_{1-\alpha/2}\sqrt{\frac{\widehat{p}\,(1-\widehat{p})}{n_t}}\right)$$

what if $\widehat{p} = 0$?

http://en.wikipedia.org/wiki/Binomial_proportion_confidence_interval

To improve the estimate

- Random Subsampling (The repeated holdout method)
- K-Fold Cross-Validation ($K = 10$ or $K = 2$ or $k = n$)
- Leave-one-out Cross-Validation ($k = 1$)
- Bootstrap

Error bars: the gaussian approximation

... and to stabilize: iterate K times - do it say $K = 10$ times

The repeated holdout method

- Holdout estimate can be made more reliable by repeating the process with different subsamples
- In each iteration, use a different random splitting
- Average the error rates on the different iterations

$$\text{mean error rate } e = \frac{1}{K}\sum_{k=1}^{K} e_k \qquad \text{variance } \widehat{\sigma}^2 = \frac{1}{K-1}\sum_{k=1}^{K}(e_k - e)^2 \quad .$$

$$e + t_{\alpha/2,K-1}\sqrt{\frac{\widehat{\sigma}^2}{K}}$$

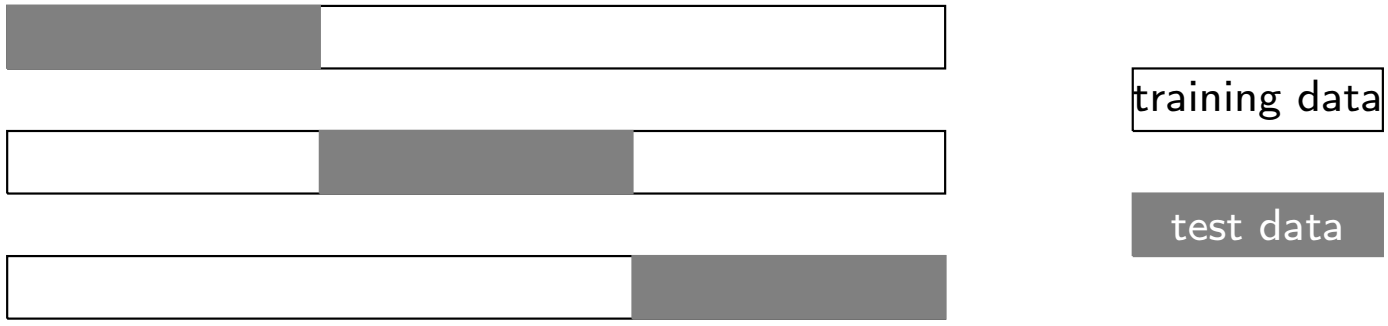
$t_{0.025,9} = 2.262$

Cross validation

Definition (Cross-validation)

A method for estimating the accuracy of an inducer by dividing the data into K mutually exclusive subsets (the "folds") of approximately equal size.

Exemple of $K = 3$-Fold Cross-Validation



How many folds are needed (K =?)

- large: small bias, large variance as well as computational time
- small: computation time reduced, small variance, large bias
- A common choice for K-Fold Cross Validation is K=5

Leave one out cross validation

Theoretical guarantees

The bootstrap

Comparing results

Two different issues

- what is the best method for my problem?
- how good is my learning algorithm?

Comparing two algorithms: Mc Nemar's test

build the confusion matrix of the two algorithms

Algo 1 / Algo 2	right	wrong
right	number of examples well classified by both	e_{01} number of examples well classified by 1 but not by 2
wrong	e_{10} number of examples missclassified by 1 but not by 2	number of examples missclassified by both

H_0: if the two algorithms are the same (we expect $e_{10} = e_{01} = \frac{e_{10}+e_{01}}{2}$)

$$\frac{(|e_{10} - e_{01}| - 1)^2}{e_{10} + e_{01}} \sim \chi_1^2$$

Beware: if $e_{10} + e_{01} < 20$ better use the sign test

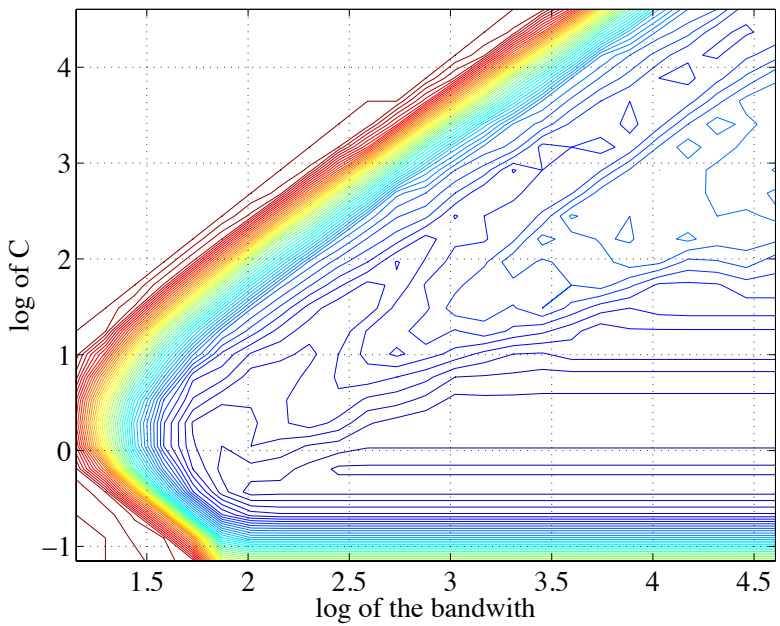
Matlab function:

`http://www.mathworks.com/matlabcentral/fileexchange/189-discrim/content/discrim/mcnemar.m`

J. L. Fleiss (1981) Statistical Methods for Rates and Proportions. Second Edition. Wiley.

Roadmap

1. Tuning hyperparameters
   - Motivation
   - Machine learning without data
   - Assessing the quality of a trained SVM
   - Model selection



"Evaluation is the key to making real progress in data mining", [Witten & Frank, 2005], p.143 (from N. Japkowicz & M. Shah ICML 2012 tutorial)

Model selection strategy

Model selection criteria attempt to find a good compromise between

- The complexity of a model
- Its prediction accuracy on the training data

1. $(Xa, ya, Xt, yt) \leftarrow split(X, y, options)$
2. $(C, b) \leftarrow tune(Xa, ya, options)$
3. $model \leftarrow train(Xa, ya, C, b, options)$
4. $error \leftarrow test(Xt, yt, C, b, options)$

Occam's Razor:
the best theory is the smallest one that describes all the facts

Model selection: the tuning function

function $(C, b) \leftarrow tune(Xa, ya, options)$

1. $(X\ell, y\ell, Xv, yv) \leftarrow split(Xa, ya, options)$
2. loop on a grid for C
3. loop on a grid for b
   1. $model \leftarrow train(X\ell, y\ell, C, b, options)$
   2. $error \leftarrow test(Xv, yv, C, b, options)$

The three sets

- Training set: a set of examples used for learning: to fit the parameters
- Validation set: a set of examples used to tune the hyper parameters
- Test set: independent instances that have played no part in formation of classifier

how to design the grids

A grid on b

A much simpler trick is to pick, say 1000 pairs (x,x') at random from your dataset, compute the distance of all such pairs and take the median, the 0.1 and the 0.9 quantile. Now pick b to be the inverse any of these three numbers.

`http://blog.smola.org/post/940859888/easy-kernel-width-choice`

A grid on C

from C_{min} to ∞

to much!

The coarse to fine strategy

1. use a large coarse grid on a few data to localize interesting values
2. fine tuning on all data in this zone

1. $(Xa, ya, Xt, yt) \leftarrow split(X, y)$
2. $(C, b) \leftarrow tune(Xa, ya, coarsegrids, smalltrainingset)$
3. $finegrids \leftarrow fit\_grid(C, b)$
4. $(C, b) \leftarrow tune(Xa, ya, finegrids, largetrainingset)$
5. $model \leftarrow train(Xa, ya, C, b, options)$
6. $error \leftarrow test(Xt, yt, C, b, options)$

The computing time is the key issue

Evaluation measures

the span bound

Bibliography

http://research.cs.tamu.edu/prism/lectures/iss/iss_l13.pdf
http://www.cs.odu.edu/~mukka/cs795sum13dm/Lecturenotes/Day3/Chapter5.pdf
http://www.cs.cmu.edu/~epxing/Class/10701-10s/Lecture/lecture8.pdf
http://www.mohakshah.com/tutorials/icml2012/Tutorial-ICML2012/Tutorial_at_ICML_2012_files/ICML2012-Tutorial.pdf