Margins, Sparsity and Perceptrons

Bob Williamson
Department of Engineering
Australian National University
Bob.Williamson@anu.edu.au

May 1, 2000

This is joint work with
Ralf Herbrich (T.U. Berlin)
Thore Graepel (T.U. Berlin)
Alex Smola (ANU)

Special thanks to Ralf Herbrich and Thore Graepel
for allowing me to present some results of
theirs for the first time.

STRUCTURE

1. Introduction
2. Approaches to Margin Results
3. Margin versus sparsity
4. Exploring version space
5. Egalitarian bound
6. Conclusions

INTRODUCTION

Input space: $\mathcal{X}$ (suppose unit ℓ_2 ball in d dimensions)

Training data:

$$Z := ((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m))$$

where $\mathbf{x}_i \in \mathcal{X}$ are generated iid according to a distribution **P** and

$$y_i = \mathrm{sgn}(\langle \mathbf{w}^*, \mathbf{x}_i \rangle).$$

Aim: Given Z, determine $\hat{\mathbf{w}}$, a "good estimate" of $\mathbf{w}^*$.

Good: By "good" we mean that the hypothesis $f_{\hat{\mathbf{w}}} : \mathbf{x} \mapsto \mathrm{sgn}(\langle \hat{\mathbf{w}}, \mathbf{x} \rangle)$ has small *generalization error*.

In other words, we wish to minimize the probability of error.

$$\begin{aligned} & \mathbf{P}\{\langle \mathbf{w}^*, \mathsf{X} \rangle \neq \langle \hat{\mathbf{w}}, \mathsf{X} \rangle\} \\ = \quad & \mathbf{P}\{\mathsf{Y} \langle \mathbf{w}, \mathsf{X} \rangle \leq 0\} \end{aligned}$$

Perceptron Algorithm (PA)

An early algorithm for finding $\hat{\mathbf{w}}$ given Z was the *Perceptron Algorithm* [Rosenblatt 1958] and was one of the earliest artificial neural networks.

$t := 0$

$\mathbf{w}_t := \mathbf{0}$

while $\exists i \in \{1, \ldots, m\}$ s.t. $y_i \langle \mathbf{w}_t, \mathbf{x}_i \rangle \leq 0$ **do**

$\mathbf{w}_{t+1} := \mathbf{w}_t + \mu y_i \mathbf{x}_i$;

$t := t + 1$;

end do;

$\hat{\mathbf{w}} := \mathbf{w}_t$;

Here $\mu > 0$ is a *step-size* parameter.

Kernel Trick — Why Linear Classifiers are so Powerful

[Bashkirov, Braverman and Muchnik (1964)]

Suppose k is a *Mercer kernel* i.e. satisfies the hypotheses of Mercer's theorem:

$$T_k f(\cdot) := \int_X k(\cdot, \mathbf{y}) f(\mathbf{y}) d\mathbf{y}$$

is positive and has continuous eigenfunctions.

Then there exists a nonlinear map Φ into a reproducing kernel Hilbert space $\mathcal{K}$ such that k computes the dot product in $\mathcal{K}$, i.e.

$$k(\mathbf{x}, \mathbf{y}) = \langle \Phi(\mathbf{x}), \Phi(\mathbf{y}) \rangle_{\mathcal{K}}.$$

The space $\mathcal{K}$ is the feature space, and is in general infinite dimensional.

Kernel Trick — 2

Replace $\langle \mathbf{w}, \mathbf{x} \rangle$ in algorithms by $k(\mathbf{x}, \mathbf{y})$.

Hypothesis becomes:

$$f(\mathbf{x}) = \sum_{i=1}^{m} \alpha_i k(\mathbf{x}_i, \mathbf{x}) \underbrace{+b}_{\text{glossed over here}} \quad ,$$

with $\alpha_i \in \mathbb{R}$, $i = 1, \ldots, m$.

The weight vector becomes

$$\mathbf{w} = \sum_{i=1}^{m} \alpha_i \phi(\mathbf{x}_i)$$

although this is never explicitly computed. (Work with $\boldsymbol{\alpha}$ always.)

Large Margin Classifiers

Solution to the problem is not unique. And indeed, if PA on Z gives a solution $\hat{\mathbf{w}}$, then in general it will give a different solution on a permutation of Z.

There is a *set* of possible solutions (the "solution set" or "version space")

Given Z (data) define

$$V := V_Z := \{\mathbf{w} : \langle \mathbf{w}, \mathbf{x}_i \rangle = y_i, (\mathbf{x}_i, y_i) \in Z\}$$

Another possibility is to seek the minimum length weight vector satisfying $\langle \mathbf{w}, \mathbf{x}_i \rangle \geq b$ for all i, where b is a positive constant called the *margin*.

The motivation behind these attempts to find a solution vector closer to the "middle" of the solution region is the intuitive belief that the resulting solution is more likely to classify new samples correctly. [Duda and Hart — 1973]

Suggests we seek a $\hat{\mathbf{w}}$ with a *large margin*.

NB: Large Margin Classifiers + Kernel Trick = Support Vector Machines

APPROACHES TO MARGIN RESULTS

Question: how to theoretically justify the idea of maximimizing the margin?

Margin of a point (Unnormalised) $\mathbf{x}_i$: $\gamma_i(\mathbf{w}) := \dfrac{y_i\langle\phi(\mathbf{x}_i),\mathbf{w}\rangle_{\mathcal{K}}}{\|\mathbf{w}\|_{\mathcal{K}}}$

Margin of a training set (Unnormalised) Z: $\gamma_Z(\mathbf{w}) := \min_{(\mathbf{x}_i,y_i)\in Z}\gamma_i(\mathbf{w})$

Expected Risk (generalisation error): $R[f_{\mathbf{w}}] := \mathbf{P}_{\mathsf{XY}}[\mathsf{Y}\langle\mathbf{w},\mathsf{X}\rangle \le 0]$

Bound $R[f_{\mathbf{w}}]$ in terms of $\gamma_Z(\mathbf{w})$

Some approaches:

- via covering numbers
- PAC Bayesian
- Cummerbund (refinement of PAC Bayesian)

via Covering Numbers

If (X, d) is a metric space; $\mathcal{F}$ a class of functions defined on X, then $\mathcal{N}(\varepsilon, \mathcal{F}, d)$ is the smallest size set U_ε such that for any $f \in \mathcal{F}$ there is a $u \in U_\varepsilon$ such that $d(f, u) < \varepsilon$.

Let $\mathbf{X}^m := (\mathbf{x}_1, \ldots, \mathbf{x}_m)$ where $\mathbf{x}_i \in \mathcal{X}$ for $i = 1, \ldots, m$. The *ε-growth function*,

$$\mathcal{N}^m(\varepsilon, \mathcal{F}) := \sup_{\mathbf{x}_1, \ldots, \mathbf{x}_m \in \mathcal{X}} \mathcal{N}(\varepsilon, \mathcal{F}, \ell_\infty^{\mathbf{X}^m}).$$

Consider the set $\mathcal{F}$ of real valued functions $\mathbf{x} \mapsto \langle \mathbf{w}, \mathbf{x} \rangle$ such that $\|\mathbf{w}\| = 1$. Suppose we are given a data set Z of size m and $\mathbf{w}$ such that $f_\mathbf{w} : \mathbf{x} \mapsto \operatorname{sgn}(\langle \mathbf{w}, \mathbf{x} \rangle)$ obtains margin $\gamma_Z(\mathbf{w}) > 0$. With probability $1 - \delta$ over a random draw of Z the expected generalization error satisfies

$$R[f_\mathbf{w}] \leq \frac{2}{m} \left(\lceil \log(\mathcal{N}^m(\gamma_Z(\mathbf{w})/2, \mathcal{F}) \rceil + \log(m) + \log(1/\delta) \right)$$

via Covering Numbers — (cont.)

The tricky bit is to bound $\log \mathcal{N}^m(\varepsilon, \mathcal{F})$.

Can do so via fat-shattering dimension or direct calculation.

Maurey: $\log \mathcal{N}^m(\varepsilon, \mathcal{F}) \leq c \log(m)/\varepsilon^2$

Can thus exploit eigenvalues associated with kernel k (see later).

Also can obtain bounds in terms of $\mathcal{N}(\varepsilon, \mathcal{F}, \ell_\infty^{\mathbf{X}^m})$ (the empirical covering numbers).

Bottom line: Rigorous justification that maximizing the margin is a good thing to do in terms of the generalization error.

Although non-asymptotic, the bounds are trivial for small sample sizes (they state a probability is less than some number much greater than 1 unless m is really quite large). :-(

PAC Bayesian Framework

[McAllester (1998); Herbrich and Graepel (1999/2000)]

Uses Bayesian analysis. Correctness of analysis does *not* depend on "correctness" of prior, although the "tightness" of the analysis does. (cf. luckiness framework)

Gibbs Classifier: Assume a prior over $\mathbf{w}$. Draw classifiers from the version space according to the posterior distribution. [Random]

Generalized Gibbs Classifier $h_{\mathrm{Gibbs}}^{H(Z)}$: Given a subset $H(Z) \subseteq V(Z)$ draw hypotheses h according to the posterior distribution restricted to $H(Z)$.

Basic PAC Bayesian Result

Any measure $\mathbf{P}_H$ (prior)

Any measure $\mathbf{P}_Z$ (data distribution)

With probability at least $1-\delta$ over a random draw of Z (according to $\mathbf{P}_Z$ of size m, for all $H(Z) \subseteq V(Z)$ such that $\mathbf{P}_H(H(Z)) > 0$,

$$R\left[h_{\text{Gibbs}}^{H(Z)}\right] \leq \frac{1}{m}\left(\ln\left(\frac{1}{\mathbf{P}_H(H(Z))}\right) + 2\ln(m) + \ln\left(\frac{1}{\delta}\right) + 1\right) \qquad \textbf{(1)}$$

Clearly, classical Gibbs algorithm minimizes the bound.

Bayes Classifier

Bayes Classifier h_{Bayes}: Given $\mathbf{x}$ predict the most likely label $h(\mathbf{x})$ when sampling h from the posterior. [Deterministic]

Generalized Bayes Classifier ${h_{\text{Bayes}}}^{H(Z)}$: Like h_{Bayes} except only hypotheses in $H(Z)$ get to "vote".

When h_{Bayes} is wrong, at least half of the classifiers in $H(Z)$ are wrong too. Thus

$$R[h_{\text{Bayes}}^{H(Z)}] \leq 2R[h_{\text{Gibbs}}^{H(Z)}] \qquad \textbf{(2)}$$

Assume prior is uniform. If $H(Z)$ is a ball with center h, then

$$R[h] = R[h_{\text{Bayes}}^{H(Z)}] \qquad \textbf{(3)}$$

("Bayes admissibility") **Amazing!**

Normalised Margin

Recall the unnormalised margin: $\gamma_Z(\mathbf{w}) = \min_{(\mathbf{x}_i, y_i) \in Z} \frac{y_i \langle \mathbf{w}, \phi(\mathbf{x}_i) \rangle_{\mathcal{K}}}{\|\mathbf{w}\|_{\mathcal{K}}}$

Define the *Normalised margin*:

$$\Gamma_Z(\boldsymbol{\alpha}) = \Gamma_Z(\mathbf{w}_{\boldsymbol{\alpha}}) = \min_{(\mathbf{x}_i, y_i) \in Z} \frac{y_i \langle \mathbf{w}_{\boldsymbol{\alpha}}, \phi(\mathbf{x}_i) \rangle_{\mathcal{K}}}{\|\mathbf{w}_{\boldsymbol{\alpha}}\|_{\mathcal{K}} \|\phi(\mathbf{x}_i)\|_{\mathcal{K}}}$$

Combining (1), (2), (3) get a bound for large margin classifiers in terms of ratio of volume of “circle” of radius Γ on surface of sphere to total surface of sphere.

Substitute a clever bound on this to obtain:

$$\frac{2}{m}\left(N \ln\left(\frac{2}{\Gamma_Z^2(\mathbf{w})}\right) + 2\ln(m) + \ln\left(\frac{1}{\delta}\right) + 2\right)$$

N

$$\frac{2}{m}\left(N \ln\left(\frac{2}{\Gamma_Z^2(\mathbf{w})}\right) + 2\ln(m) + \ln\left(\frac{1}{\delta}\right) + 2\right)$$

This bound is very nice (tightest yet for linear classifiers) . But when using a kernel, $N = m$: problem!

Know from other analyses that "effective number of dimensions" can be far less than m.

Idea: truncate empirical eigenvalues. Get small N, and bound error due to truncation.

Can get a transductive result. Complicated to state (not done here).

Difficulty for inductive result is feature space depends on $\mathbf{x}_1, \ldots, \mathbf{x}_m$.

Conjecture: still the best route for SVM bounds.

Exploiting the Input Distribution

PAC-Bayesian Margin bound uses uniform prior. In practice most hypothesis (drawn uniformly over the sphere) classify all of the training data to the same class.

Question: by ruling out such hypotheses (those that assign all points to the same class) can we tighten the bound further?

Executive summary: Yes. Key point: Knowing a region *a priori* where the data lies can conclude there is only a subset of parameter space that can classify the data non-constantly.

The Cummerbund

Given a convex set $C \subset \ell_2^N$, the *Cummerbund of* C is

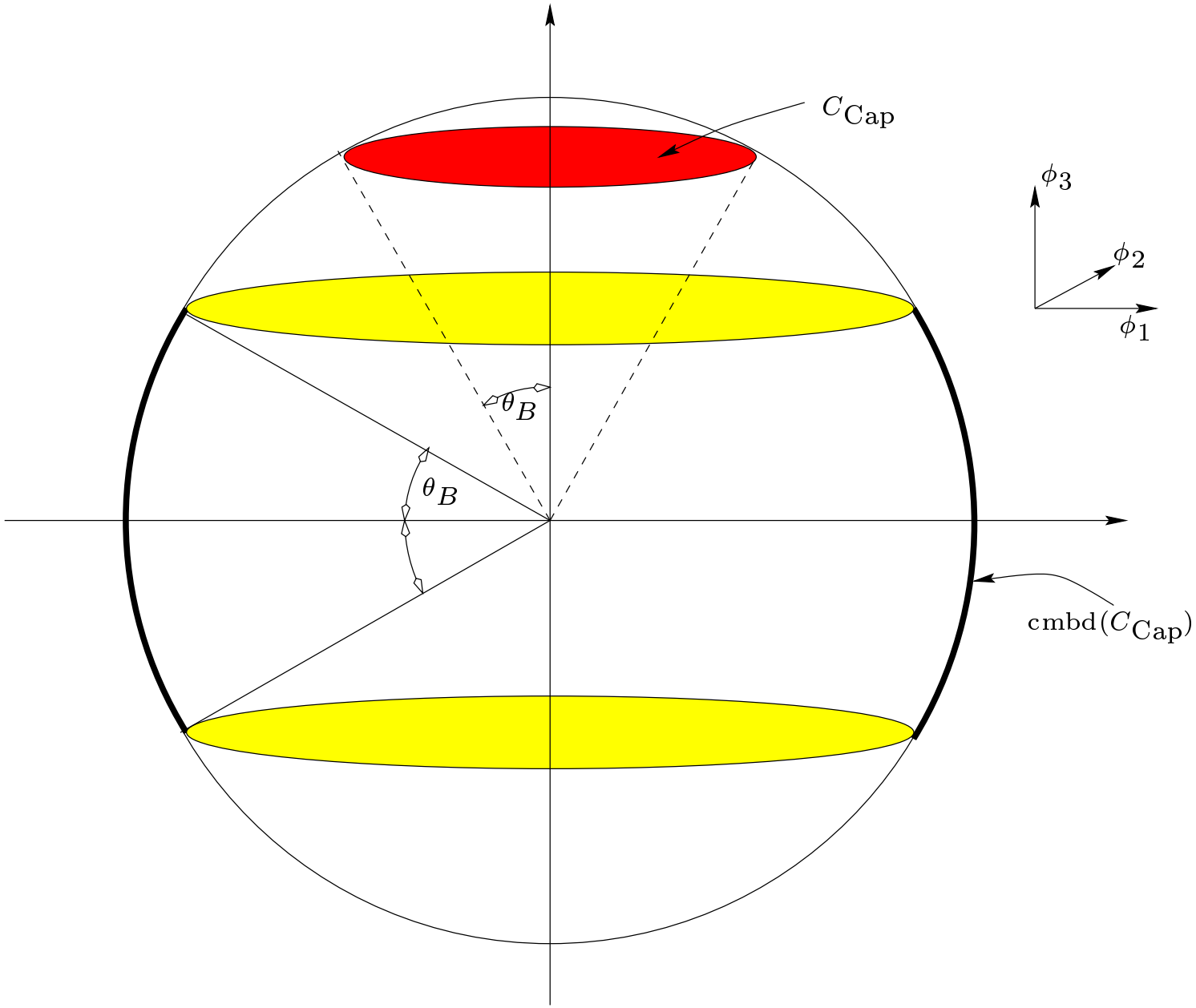
$$\mathrm{cmbd}(C) := \{\mathbf{w} \in S_N : \exists \mathbf{x} \in C, \langle \mathbf{w}, \mathbf{x} \rangle = 0\}$$

Thus $\mathrm{cmbd}(C)$ is the set of weight vectors that non-constantly classify points $\mathbf{x} \in C$.

PAC Bayesian Margin bound becomes

$$\frac{2}{m}\left(N \ln\left(\frac{\mathrm{vol}(\mathrm{cmbd}(C))}{\mathrm{vol}(S_N)}\frac{2}{\Gamma_Z^2(\mathbf{w})}\right) + 2\ln(m) + \ln\left(\frac{1}{\delta}\right) + 2\right)$$

Example: Suppose all the training data is known (a priori) to live in a "polar cap" C_{Cap} subtended by angle θ_B. Can explicitly determine $\mathrm{cmbd}(C_{\mathrm{Cap}})$ (the tropics).


C_{Cap}
ϕ_3
ϕ_2
ϕ_1
θ_B
θ_B
$\mathrm{cmbd}(C_{\mathrm{Cap}})$

Example of Cummerbund

Let $c = \cos(\theta_B)$. A geometric argument gives

$$\frac{\mathrm{vol}(\mathrm{cmbd}(C_{\mathrm{Cap}}))}{\mathrm{vol}(S_N)} \leq 1 - \left(1 - \sqrt{1 - c^2}\right)^N := F$$

$\theta_B = \pi/2 \Rightarrow c = 0 \Rightarrow F = 1$ $\theta_B = 0 \Rightarrow c = 1 \Rightarrow F = 0$

Bottom line: possible to incorporate additional prior knowledge.

MARGIN VERSUS SPARSITY

Two explanations of the power of Support Vector machines [Vapnik]

1. SVMs provide a *sparse* solution. Bounds on expected generalization error in terms of sparseness of solution.

2. SVMs maximize the margin and large margins are good (as considered before). High probability bounds on the generalization error in terms of the size of the margin

Which to believe?

Luckiness framework (1998): either (your choice)
More recently (Herbrich and Shawe-Taylor): exploit both
Present talk: both, but sparsity is in some sense more fundamental?

- Dual Perceptron: Sparsity of solution in terms of margin.
- Geometric argument: Large margin implies there exists a sparse solution.

Dual Perceptron

[Aizerman, Braverman and Rozonoer (1964)]

$t := 0;$

$\boldsymbol{\alpha}_t := \mathbf{0}$ # Differences in red

while $\exists i \in \{1, \ldots, m\}$ s.t. $y_i \langle \mathbf{w}_{\boldsymbol{\alpha}_t}, \phi(\mathbf{x}_i) \rangle_{\mathcal{K}} \leq 0$ **do**

$\mathbf{w}_{\boldsymbol{\alpha}_{t+1}} := \mathbf{w}_{\boldsymbol{\alpha}_t} + \mu y_i \frac{\phi(\mathbf{x}_i)}{\|\phi(\mathbf{x}_i)\|_{\mathcal{K}}};$

$\boldsymbol{\alpha}_{i,t+1} = \boldsymbol{\alpha}_{i,t} + \frac{\mu y_i}{\sqrt{k(\mathbf{x}_i, \mathbf{x}_i)}};$ # Implement this

$t := t + 1;$

end do;

$\hat{\mathbf{w}} := \sum_{i=1}^{m} \boldsymbol{\alpha}_{i,t} \phi(\mathbf{x}_i)$ # Not explictly calculated

Recent interest due to Freund and Schapire (and others).

Dual Perceptron — Convergence

Recall the *Normalised margin*:

$$\Gamma_Z(\boldsymbol{\alpha}) = \Gamma_Z(\mathbf{w}_{\boldsymbol{\alpha}}) = \min_{(\mathbf{x}_i, y_i) \in Z} \frac{y_i \langle \mathbf{w}_{\boldsymbol{\alpha}}, \phi(\mathbf{x}_i) \rangle_{\mathcal{K}}}{\|\mathbf{w}_{\boldsymbol{\alpha}}\|_{\mathcal{K}} \|\phi(\mathbf{x}_i)\|_{\mathcal{K}}}$$

New version of Novikoff's theorem [Graepel and Herbrich]:

Let $Z = (X, Y)$ be a training set of size m and suppose k is a Mercer kernel. Suppose there exists $\boldsymbol{\alpha}^* \in \mathbb{R}^m$ such that $\Gamma_Z(\boldsymbol{\alpha}^*) > 0$. Then the number of mistakes made by the kernel perceptron algorithm on Z is at most

$$\left(\frac{1}{\Gamma_Z(\boldsymbol{\alpha}^*)} \right)^2 .$$

Dual Perceptron — Convergence (cont.)

Observe that if $\mu = 1$ then this implies $\|\boldsymbol{\alpha}_T\|_1 \leq \Gamma_Z(\boldsymbol{\alpha}^*)^{-2}$ (T is last step).

If in addition $k(\mathbf{x}_i, \mathbf{x}_i) = 1$ (e.g. RBF kernels) the components of $\boldsymbol{\alpha}_T$ are integers and so $\|\boldsymbol{\alpha}_T\|_0 \leq \|\boldsymbol{\alpha}_T\|_1$.

$\|\boldsymbol{\alpha}_T\|_0$ is the number of nonzero components of $\boldsymbol{\alpha}_T$.

Thus the previous result bounds the sparsity of $\boldsymbol{\alpha}_T$.

Now the *Compression Lemma* [Littlestone and Warmuth] can be used to obtain generalization error bound for Dual Perceptron [Graepel, Herbrich and Shawe-Taylor]:

If the dual perceptron converges to a solution $\boldsymbol{\alpha}_T$ then its generalization error is less than

$$\frac{1}{m - \|\boldsymbol{\alpha}_T\|_0} \left(\ln \binom{m}{\|\boldsymbol{\alpha}_T\|_0} + \ln(m) + \ln\left(\frac{1}{\delta}\right) \right).$$

Dual Perceptron — Generalization Error Bound

Substitute bound on $\|\boldsymbol{\alpha}\|_0$ in terms of margin to obtain:

For any distribution **P**, with probability at least $1-\delta$ over the random draw from $\mathbf{P}^m$ of a training set Z of size m, if there exists a vector $\boldsymbol{\alpha}^*$ such that $\Gamma_Z(\boldsymbol{\alpha}^*) > \sqrt{\frac{1}{m}}$ then the generalization error of the classifier $\boldsymbol{\alpha}_T$ found by the dual perceptron algorithm is less than

$$\frac{1}{m - \Gamma_Z^{-2}(\boldsymbol{\alpha}^*)} \left(\ln \binom{m}{\Gamma_Z^{-2}(\boldsymbol{\alpha}^*)} + \ln(m) + \ln\left(\frac{1}{\delta}\right) \right).$$

Remarkable: *mere existence* of a large margin classifier $\boldsymbol{\alpha}^*$ suffices to ensure the good generalization performance of a *different* classifier $\boldsymbol{\alpha}_T$.

Dual Perceptron — Empirical Results

Freund and Schapire's empirical results for dual perceptron and SVMs.

NIST dataset. Polynomial kernel of degree 4. $m = 60000$. Generalisation error (in %) and [Graepel, Herbrich and Shawe-Taylor] bound (in % with $\delta = 0.05$).

Digit	0	1	2	3	4	5	6	7	8	9
DualPerc	0.2	0.2	0.4	0.4	0.4	0.4	0.4	0.5	0.6	0.7
$\|\boldsymbol{\alpha}\|_0$	740	643	1168	1152	1078	1277	823	1103	1856	1920
Mistakes	844	843	1345	1811	1222	1497	960	1323	2326	2367
Bound	6.7	6.0	9.8	12.0	9.2	10.5	7.4	9.4	14.3	14.6
SVM	0.2	0.1	0.4	0.4	0.4	0.5	0.3	0.4	0.5	0.6
$\|\boldsymbol{\alpha}\|_0$	1379	989	1958	1900	1224	2024	1527	2064	2332	2765
Bound	11.2	8.6	14.9	14.5	10.2	15.3	12.2	15.5	17.1	19.6

Number of mistakes a good estimate of $\|\boldsymbol{\alpha}\|_0$.

SVM works (slightly) better but perceptron bound is better.

Dual perceptron performs well when SVM does.

A Direct Geometric Connection

Question: A more direct connection between large margin and large sparsity?

Consider a fixed feature space $\mathcal{K}$ of dimensionality N with an associated feature space map $\phi\colon \mathcal{X} \to \mathcal{K}$. For any $\lambda \in [0,1]$ and all data sets $Z = (X,Y) \in (\mathcal{X} \times \{-1,1\})^m$ of size m, **if** there exists $\boldsymbol{\alpha}^*$ such that $\|\mathbf{w}_{\boldsymbol{\alpha}^*}\| = 1$ and

$$\Gamma_Z(\boldsymbol{\alpha}^*) > \sqrt{1-\lambda}$$

then there exists $\boldsymbol{\alpha} \in \mathbb{R}^m$ such that $\Gamma_Z(\boldsymbol{\alpha}) > 0$ and

$$\|\mathbf{w}_{\boldsymbol{\alpha}}\|_0 \leq \lceil \lambda N \rceil.$$

Here λ is the *relative sparseness* (proportion of non-zero components) of $\mathbf{w}$ (not $\boldsymbol{\alpha}$).

A Direct Geometric Connection (Cont.)

Recall

$$\mathbf{w} = \sum_{i=1}^{m} \alpha_i \phi(\mathbf{x}_i)$$

If $\phi(\mathbf{x}_i) = e_i$ (canonical basis) (or equivalently if k is linear kernel so $\phi\colon \mathbf{x} \mapsto \mathbf{x}$ and $(\mathbf{x}_i)_i$ are orthonormal), then $\mathbf{w}_{\boldsymbol{\alpha}} = \boldsymbol{\alpha}$.

In that case get

$$\exists \boldsymbol{\alpha}^*, \Gamma_Z(\boldsymbol{\alpha}^*) > \sqrt{1-\lambda} \quad \Rightarrow \quad \exists \boldsymbol{\alpha}, \|\boldsymbol{\alpha}\|_0 \leq \lceil \lambda N \rceil$$

Conjecture: This result holds in general.

EXPLORING VERSION SPACE

Have seen several theoretical justification for maximum margin hyperplane.

Given a prior (e.g. uniform) could use the Bayes point (center of mass of $V(Z)$. (Hard to find algorithmically)

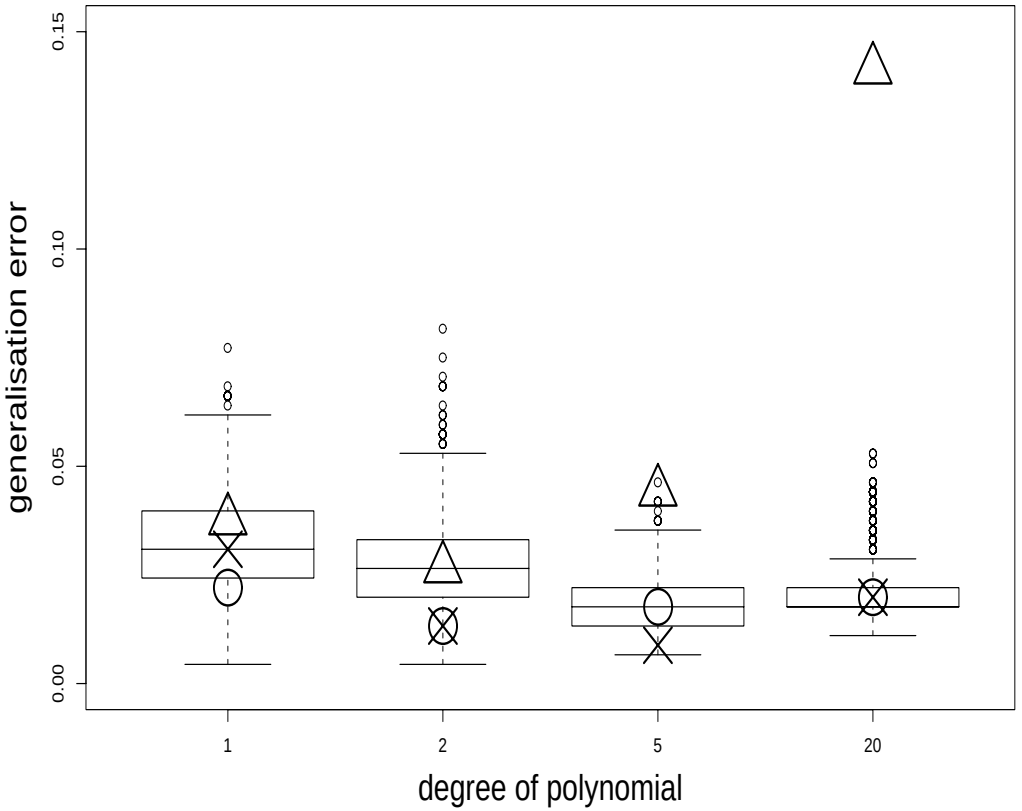
On the other hand perceptron algorithm “just” finds its way into version space (no guarantee it will be in the “middle”.

However there are good generalization error bounds for it.

Also performs almost as well as Maxmimum margin.

Question: How much variation is there in generalization performance for classifiers from all over version space?

Exploring Version Space — Empirically (1)

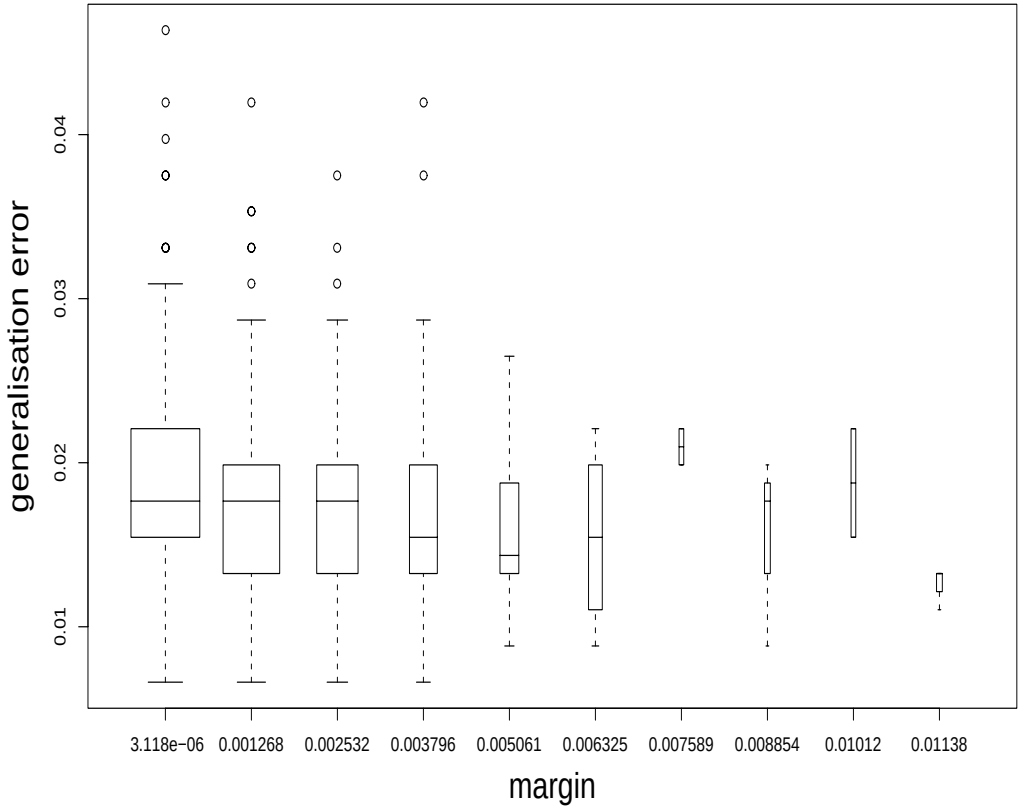


Kernel: $k(\mathbf{x}_i, \mathbf{x}_j) = (\langle \mathbf{x}_i, \mathbf{x}_j \rangle + 1)^p$.

Run "Kernel Gibbs Sampler" 1000 times on MNIST dataset (digits "1" and "2") ($m = 118$, test-set size 453).

SVM $\triangle$, SVM normalised in $\mathcal{K}$ $\times$, Bayes Point Machine: $\circ$.

Exploring Version Space — Empirically (2)



Generalisation error for different attained margins for $p = 5$.

Width of box proportional to number of samples it was based upon.

EGALITARIAN BOUND

Starting point: Explain previous results!

Puzzle: Maximum margin hyperplane is in the "middle" of version space. Kernel Perceptron solution near the edge. Generalization performance not that different!

Question: How much variation in generalization performance is there as one moves h^* over $V(Z)$.

Approach: Draw a straight-forward conclusion from PAC Bayesian Bound.

Recall

$$\varepsilon_{\text{Gibbs}}(m, \mathbf{U}_H, Z, \delta) := R\left[h_{\text{Gibbs}}^{V(Z)}\right] \leq \frac{1}{m}\left(\ln\left(\frac{1}{\mathbf{U}_H(V(Z))}\right) + 2\ln(m) + \ln\left(\frac{1}{\delta}\right) + 1\right)$$

Egalitarian Bound

For all measures $\mathbf{P}_Z$, with probability at least $1 - \delta$ over a random draw of the training set Z of size m, for all $\eta > 1$, at least a fraction of $1 - \frac{1}{\eta}$ of the classifiers in version space $V(Z)$ have a generalisation error less than

$$\eta \cdot \varepsilon_{\text{Gibbs}}(m, \mathbf{U}_H, Z, \delta).$$

Interpretation: Consider $\eta = 2$. Recall

$$R[h^{V(Z)}_{\text{Bayes}}] \leq 2R[h^{V(Z)}_{\text{Gibbs}}]$$

Above results says *whenever the Bayes solution has a small generalization error bound at least half of the consistent classifiers have the same bound*.

Of course, still need to be able to *find* one of these good solutions.

Egalitarian Bound — Cont.

Also recall Max Margin bound:

$$R[h_{\Gamma(Z)}] \leq 2R[h^{\Gamma-\text{ball}}_{\text{Gibbs}}]$$

Since $R[h^{\Gamma-\text{ball}}_{\text{Gibbs}}] \geq R[h^{V(Z)}_{\text{Gibbs}}]$, with $\eta = 2$ result also says

whenever the Maximum margin solution has a small generalization error bound, at least half of the consistent classifiers have the same bound.

Note in high dimensional spaces, most of the "volume" is near the edges.

Conclusion: Additional support for kernel perceptron, approximate BPM etc: Maximum margin hyperplane losing its "special status"?

CONCLUSIONS

- Perceptrons of enduring interest because of kernel trick.
- Refined PAC Bayesian analysis of generalization performance of large margin hyperplanes (Cummerbund).
- Kernel Perceptron obtains a sparse solution whenever a large margin solution exists.
- Geometrical argument that large margin implies sparsity.
- Tight(er) bounds on generalization performance via sparsity.
- Sparsity based bounds have the advantage that do not seem to have to exploit eigenvalues to get good bounds (cf. Margin bounds).
- Maximal margin is *one* choice of a point in version space. Bayes point is better, but hard to find.
- Egalitarian bound shows many choices within version space do as well as Gibbs. Max Margin not so special?