Reliability of Cross-Validation for SVMs in High-Dimensional, Low Sample Size Scenarios

Sascha Klement, Amir Madany Mamlouk, and Thomas Martinetz

Institute for Neuro- and Bioinformatics, University of Lübeck

Abstract. A Support-Vector-Machine (SVM) learns for given 2-class-data a classifier that tries to achieve good generalisation by maximising the minimal margin between the two classes. The performance can be evaluated using cross-validation testing strategies. But in case of low sample size data, high dimensionality might lead to strong side-effects that can significantly bias the estimated performance of the classifier. On simulated data, we illustrate the effects of high dimensionality for cross-validation of both hard- and soft-margin SVMs. Based on the theoretical proofs towards infinity we derive heuristics that can be easily used to validate whether or not given data sets are subject to these constraints.

1 Introduction

Learning from examples in cases where many degrees of freedom but only few examples are available is commonly considered a challenging problem. Due to massively parallel data acquisition systems, such as microarray gene analysis [1,2], it easily happens that there are very few data points described by thousands of features. In such cases, theoretical and practical issues such as generalisation bounds, run-time, or memory-footprint considerations require sophisticated validation methods to measure the performance of machine learning algorithms in high-dimensional feature spaces.

Two commonly used methods in this context are support vector machines (SVM, [3,4]) as a classification system and cross-validation (CV, [5]) as a validation tool. Both methods are closely connected. Typically, when using the SVM there is a tendency to increase the data dimensionality as the classification problem is simplified in higher dimensions; on the other hand, CV is the method of choice in scenarios with relatively few data points compared to the dimensionality.

There is a well-known effect that if dimensionality is increased towards infinity, a finite set of points will lose more and more of its spatial topology. In the limit, the points will be located on the vertices of a regular simplex [6], i.e. all samples have nearly the same distances to the origin as well as among each other, and they are pairwise orthogonal. These properties were shown for multivariate standard normal distributions with zero mean and identity covariance matrix but hold under much weaker assumptions [7].

V. Kurkova, R. Neruda, and J. Koutnik (Eds.): ICANN 2008, Part I, LNCS 5163, pp. 41–50, 2008.

Usually, the dimensionality will be finite, but we will show that even comparatively low dimensional data will behave as if being infinitely dimensional. So, especially for low sample size data, *infinity* is rather *small*.

First, we show that the leave-one-out CV error for hard-margin SVMs will approach 1 in high-dimensional feature spaces for equal-sized classes drawn from the same distribution – despite the expected error rate of 0.5, which would be the outcome for the same setting in low dimensions.

This first observation is generalised to two classes drawn from different distributions. Hall [6] showed that whenever these classes are too close together, a hard-margin SVM will vote along the majority rule alone for a dimension towards infinity. We will show empirically that this will occur even in quite finite dimensions when the given sample size is small.

Finally, we show the soft-margin approach to make things even worse. One might think that only simple hard-margin SVMs are prone to severe overfitting. Soft-margin SVMs increase the margin to reduce the fat-shattering dimension and should therefore reduce overfitting by allowing training errors. Unfortunately, this does not increase the generalisation performance, again due to the counterintuitive geometric properties of only few samples in high-dimensional space and the asymmetries of a resampling scheme such as leave-one-out cross-validation. In the soft-margin case infinity becomes even *smaller*.

It should be emphasised that especially dealing with high-dimensional but small sample size data leads to various counterintuitive and unfamiliar side-effects, which can significantly impact training and validation. In tasks such as the analysis of genetic microarrays, practical and financial issues strictly limit the maximum sample size but all the same rely on the analysis of high-dimensional content. Therefore, we want to elucidate the constraints on dimensionality, sample size, and class distribution, which might help to make training with SVMs still feasible in such scenarios.

2 Leave-One-Out and Hard-Margin SVM

Leave-one-out cross-validation is commonly used to estimate the generalisation performance and to fine-tune training parameters for various machine learning algorithms. The computational complexity limits the usage to small sample size data, but there it is commonly regarded to give good approximations of the true generalisation performance.

Still, it will fail in certain scenarios such as in the following example. Assume a balanced two class random dataset, i.e. samples drawn from an arbitrary distribution with randomly assigned class labels. The best classifier for completely random datasets is simply the majority voting rule [5]. Unfortunately, leave-one-out cross-validation will indicate very poor performance, since the original balanced dataset becomes unbalanced in each and every validation step. As the left-out pattern reduces the size of one class, the majority classifier will always vote for the other larger class, but the left-out pattern belongs to the smaller class. Thus, the classifier will always vote wrong. This behaviour is independent of dimensionality or training set size but requires the prior knowledge of dealing with completely random datasets.

In general, we do not have this knowledge and want to apply a generic classification framework such as linear SVMs for separation. To examine this framework in detail, we define a very simple unlearnable scenario with the training set

$$\mathcal{D} = \{\boldsymbol{x}_i, y_i\}_{i=1}^{n} \tag{1}$$

consisting of feature vectors $\boldsymbol{x}_i \in \mathbb{R}^d$ where each entry is drawn from the standard normal distribution and corresponding class labels $y_i \in \{-1, +1\}$. Without loss of generality we set

$$y_1 = \ldots = y_{\frac{n}{2}} = -1 \quad \text{and} \tag{2}$$

$$y_{\frac{n}{2}+1} = \ldots = y_n = +1 \ , \tag{3}$$

i.e. the training set represents two equal-sized classes drawn from the same distribution. For high-dimensional low sample size data $d \gg n$ holds, and therefore a separating hyperplane exists in general, except for cases with three or more collinear data points having alternating class labels.

We used leave-one-out cross validation to approximate the generalisation performance on this degenerated dataset by training a linear SVM n times using the MinOver algorithm [4], each time on a different subset of size $n - 1$. The resulting classification functions $f_i(\boldsymbol{x})$ are then used to classify the remaining pattern and the leave-one-out error E is determined by

$$E = \frac{1}{2n} \sum_{i=1}^{n} |f_i(\boldsymbol{x}_i) - y_i| \ .$$

This procedure was repeated 100 times for each $n \in \{4, 6, \ldots, 60\}$ and fixed $d = 1000$ (see Fig. 1, left) as well as for $n = 20$ and logarithmic-spaced

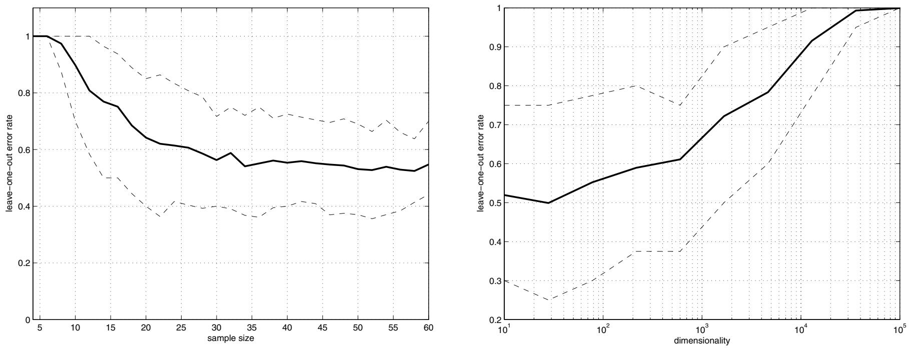

Fig. 1. The leave-one-out error rate of a hard-margin SVM tends to 1 for fixed dimensionality d and decreasing sample size n (left) as well as for fixed sample size and increasing dimensionality (right). Solid lines depict the mean of 100 trials, while dashed lines mark the 5th and 95th percentiles.

$d \in [10, 10^5]$ (see Fig. 1, right). Obviously, E depends on the training set size and tends to 1 for $n \to 4$ and $d \to \infty$. For larger training sets E tends to 50%, which is what we would intuitively assume. In order to explain error rates of close to 1 for small sample sizes, we switch from finite dimensionality to the remarkable case of finite datasets and infinite dimensionality. We will give a proof for this observation only using the geometric properties of high-dimensional small sample size data and simple vector algebra.

Theorem 1. *For any dataset $\mathcal{D} = \{\boldsymbol{x}_i, y_i\}_{i=1}^n$ with $\boldsymbol{x}_i \in \mathbb{R}^d$ drawn from the multivariate standard normal distribution, $y_1 = \ldots = y_{\frac{n}{2}} = -1$ and $y_{\frac{n}{2}+1} = \ldots = y_n = +1$ with n fixed the leave-one-out error-rate of a hard-margin SVM is 1 for $d \to \infty$.*

Proof. For $d \to \infty$ all $\boldsymbol{x}_i \in \mathcal{D}$ will lie on the vertices of a regular n-simplex as well as all pairwise angles will be orthogonal [6]. The total variability of $\mathcal{D}$ is provided in the rotation of this simplex. Without loss of generality we set $\boldsymbol{x}_i = \sqrt{d}\,\boldsymbol{e}_i$, so that for $d \to \infty$

$$\begin{aligned} ||\boldsymbol{x}_i||_2 &= \sqrt{d} \qquad \forall\, i \\ ||\boldsymbol{x}_i - \boldsymbol{x}_j||_2 &= ||\sqrt{d}\,\boldsymbol{e}_i - \sqrt{d}\,\boldsymbol{e}_j||_2 = \sqrt{2\,d} \qquad \forall\, i \neq j \\ \boldsymbol{x}_i^T \boldsymbol{x}_j &= d\,\boldsymbol{e}_i^T \boldsymbol{e}_j = 0 \qquad \forall\, i \neq j \ . \end{aligned}$$

Again without loss of generality we select $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n-1}$ for training. Now, we can analytically determine the maximum margin classifier

$$\begin{aligned} f(\mathbf{x}) &= \operatorname{sgn}\left(\boldsymbol{w}^T \boldsymbol{x} + b\right) \\ \text{that minimises} \quad & \boldsymbol{w}^T \boldsymbol{w} \\ \text{subject to} \quad & y_i \left(\boldsymbol{w}^T \boldsymbol{x}_i + b\right) \geq 1 \qquad \forall\, i \end{aligned}$$

with simple vector algebra. Since all samples are pairwise orthogonal also the centroids of both classes are orthogonal. Thus, the separating hyperplane with maximum margin is orthogonal to the straight line through the centroids (see Fig. 2), i.e.

$$\boldsymbol{w} = \bar{\boldsymbol{x}}^+ - \bar{\boldsymbol{x}}^- \quad \text{with} \quad \bar{\boldsymbol{x}}^+ = \frac{1}{\frac{n}{2}} \sum_{i=1}^{\frac{n}{2}} \sqrt{d}\,\boldsymbol{e}_i \quad \text{and} \quad \bar{\boldsymbol{x}}^- = \frac{1}{\frac{n}{2} - 1} \sum_{i=\frac{n}{2}+1}^{n-1} \sqrt{d}\,\boldsymbol{e}_i \ .$$

With $\boldsymbol{p}$ as the centre of this line we get the bias through

$$b = -\boldsymbol{w}^T \boldsymbol{p} = -\left(\bar{\boldsymbol{x}}^+ - \bar{\boldsymbol{x}}^-\right)^T \frac{1}{2} \left(\bar{\boldsymbol{x}}^+ + \bar{\boldsymbol{x}}^-\right) = \frac{2\,d}{n\,(n-2)} \ .$$

The left-out data point $\boldsymbol{x}_n$ gets misclassified, because

$$f(\boldsymbol{x}_n) = \boldsymbol{w}^T \sqrt{d}\,\boldsymbol{e}_n + b = b > 0 \ .$$

This will be the case in each and every validation step, so the total leave-one-out classification error is 1. □

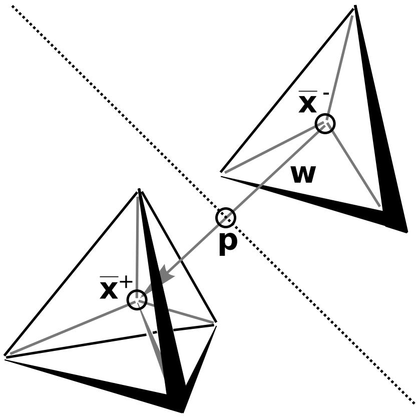

Fig. 2. Geometric sample configuration used in the proof of Theorem 1. Here, the case of 7 samples is visualised. The samples from the larger class form a 4-simplex while those from the smaller class form a 3-simplex. The centroids of both simplices are denoted by $\bar{\boldsymbol{x}}^+$ and $\bar{\boldsymbol{x}}^-$ respectively. The separating hyperplane with maximum margin has the normal vector $\boldsymbol{w}$ and contains the point $\boldsymbol{p}$.

As dimensionality decreases, the data points differ more and more from the vertices of a regular simplex and are no longer orthogonal. So the conditions of the above proof are only approximately fulfilled. Therefore, probability increases for data points to be correctly classified, and the leave-one-out error-rate will be less than 1 and will converge to the intuitive error-rate of 0.5 for $n \to \infty$.

3 Practical Bounds for Critical Scenarios

So far, we examined the leave-one-out error in the limit for $d \to \infty$ which seems to be unrealistic in practical cases. But for high dimensional low sample size data infinity is rather *small* as our next experiment will stress. We show empirically which size dimensionality needs to have to behave as if it were infinite. Again, we sampled two classes independently and identically distributed from the standard normal distribution and trained support vector machines for $n \in \{4, 6, \ldots, 40\}$ and logarithmic-spaced $d \in [10, 10^5]$. For each configuration the procedure was repeated 10 times and the mean leave-one-out error was determined.

The colour-coded results are shown in Fig. 3. For any fixed number of data points n (e.g. $n = 20$) the leave-one-out error reaches 1 for a specific dimensionality (in this case at about $d = 10000$). Due to the probabilistic behaviour of the dataset, a precise borderline does not exist, but the tendency is obvious. We illustrate this by means of linear, quadratic and exponential extrapolation derived from the approximate border points of $n \in \{4, 6, 8\}$. A linear border is obviously too low while exponential behaviour is too steep – a slightly super-quadratic tendency is most promising. However, this heuristic suits only for illustrating

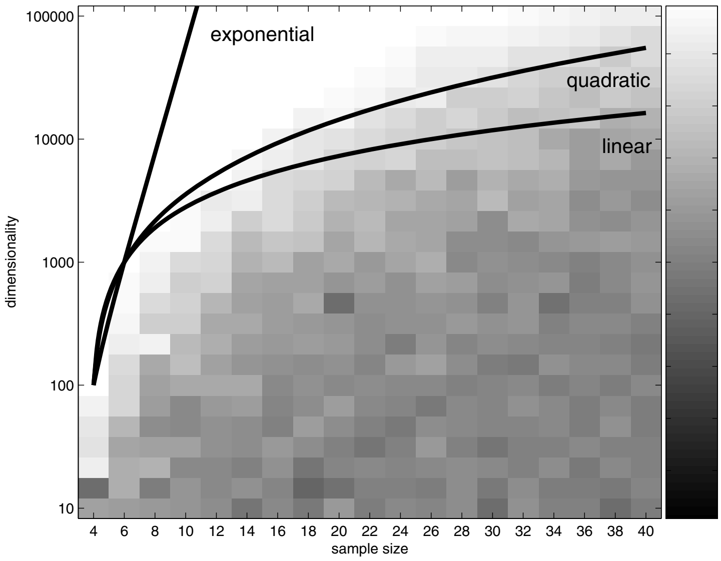

Fig. 3. Colour-coded error-rates depending on sample size and dimensionality. White colour indicates an error-rate of 1, i.e. the corresponding parameter sets behave as if they had infinite dimensionality. Additionally, three candidates for the border function are shown.

the asymptotic behaviour and should not be taken as an exact mathematical coherence.

3.1 Real Two-Class Scenarios

Now, we will generalise the random scenario from Sect. 2 to the case where both classes are drawn from different distributions. Assume the samples of the two classes to be distributed as $\boldsymbol{\mathcal{X}} = \left(\mathcal{X}^{(1)}, \ldots, \mathcal{X}^{(d)}\right)^T$ and $\boldsymbol{\mathcal{Y}} = \left(\mathcal{Y}^{(1)}, \ldots, \mathcal{Y}^{(d)}\right)^T$. With $d \to \infty$ the following conditions shall hold:

$$\sigma^2 = \lim_{d\to\infty} \frac{1}{d} \sum_{i=1}^{d} \mathrm{var}(\mathcal{X}^{(i)})$$

$$\tau^2 = \lim_{d\to\infty} \frac{1}{d} \sum_{i=1}^{d} \mathrm{var}(\mathcal{Y}^{(i)})$$

$$\mu^2 = \lim_{d\to\infty} \frac{1}{d} \sum_{i=1}^{d} \left(E(\mathcal{X}^{(i)}) - E(\mathcal{Y}^{(i)})\right)^2 \quad .$$

Any Gaussian or rectangular distributions in d dimensions fulfil these conditions. Let k and l be the number of data points of the data sets X and Y drawn from $\boldsymbol{\mathcal{X}}$ and $\boldsymbol{\mathcal{Y}}$, respectively. If we then train a hard-margin SVM on these data point this leads to [6]:

Theorem 2. *Assume that $\frac{\sigma^2}{k} \geq \frac{\tau^2}{l}$ (otherwise interchange X and Y).*

If $\mu^2 > \frac{\sigma^2}{k} - \frac{\tau^2}{l}$, then the probability of a new datum from either population to be correctly classified by the SVM converges to 1 as $d \to \infty$. Otherwise any new sample will be classified as belonging to class Y as $d \to \infty$.

Again, we want to visualise how *small* infinite dimensionality in this case is. We sampled two equal-sized classes each from the multivariate standard normal distribution so that $\sigma^2 = \tau^2 = 1$ with the total number of data samples $n \in \{4, 6, \ldots, 40\}$ and logarithmic-spaced $d \in [10, 10^5]$. We illustrate exemplarily the dependency of sample size and leave-one-out error for an interclass distance of $\mu^2 = \frac{1}{30}$. The mean colour-coded results of 10 independent runs are shown in Fig. 4. As stated in Theorem 2, a certain threshold exists – separating convergence to zero error from convergence to an error rate of one. Obviously, the threshold corresponds in this case to a total sample size of 12, i.e. 5 samples from one class are trained vs. 6 samples from the other class within each validation step. Then, the proposed ratio exceeds the critical interclass distance at

$$\frac{\sigma^2}{k} - \frac{\tau^2}{l} = \frac{1}{5} - \frac{1}{6} = \frac{1}{30} = \mu^2 \ ,$$

which is perfectly the initially chosen value.

As a consequence of Theorem 2, it follows that two classes having different means in one dimension and having the same mean in all other dimension will not be separable, since μ is close to zero. So, the leave-one-out error rate will *always* converge to exactly 1 for any sample size as dimensionality goes to infinity.

3.2 Soft Margin and the Fat-Shattering Dimension

Hard-margin SVMs are prone to overfitting since one single outlier will strongly affect the separating hyperplane and reduce generalisation performance.

In the simplest case of two equal-sized classes drawn from the same distribution – defined as in (1)–(3) – we can explicitly derive an upper bound for the fat-shattering dimension *fat* according to [8] by

$$fat(\gamma) \leq \min\left\{ \left\lceil \frac{R^2}{\gamma^2} \right\rceil, d \right\} + 1 \tag{4}$$

where R is the radius of the smallest enclosing sphere containing all samples and γ denotes the margin of separation. For randomly distributed datasets with $d \to \infty$ as used before the margin derives to

$$\gamma = \frac{1}{2}||\boldsymbol{w}|| = \sqrt{\frac{d\,(n-1)}{n\,(n-2)}} \ .$$

Using (4) and $R = \sqrt{d}$ the fat-shattering dimension is upper bounded by

$$fat(\gamma) \leq \min\left\{ \left\lceil \frac{d\,n\,(n-2)}{d\,(n-1)} \right\rceil, d \right\} + 1 = \min\{n-1, d\} + 1 = n \ .$$

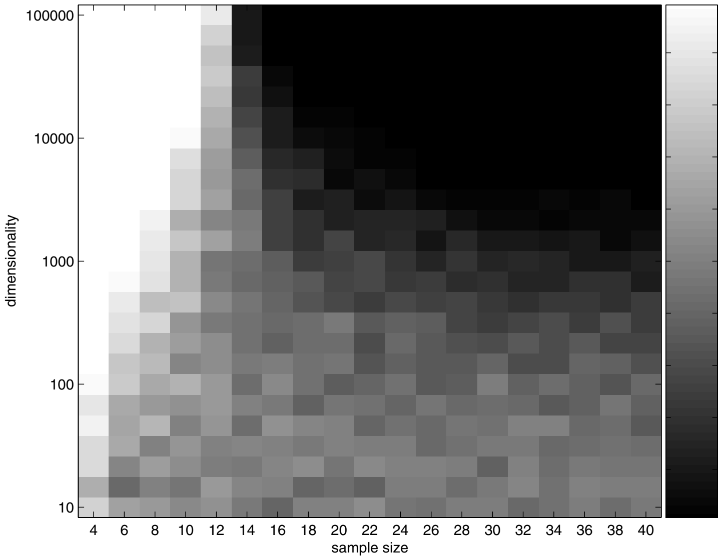

Fig. 4. Colour-coded error-rates depending on sample size and dimensionality for $\mu^2 = \frac{1}{30}$. White indicates an error-rate of 1, black an error rate of 0. The parameter sets corresponding to black or white behave as if they had infinite dimensionality.

Here the bounds $(n-2) < \frac{n\,(n-2)}{n-1} < (n-1)$ are used. Since $d \gg n$ and all samples are pairwise orthogonal, $\mathit{fat}(\gamma)$ is also lower bounded by n. Thus, in the hard-margin case the fat-shattering dimension indicates that no generalisation is possible, as expected. Considering a 2-norm soft-margin SVM, the following kernel has to be used instead of the dot product in the dual representation [3]:

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \Phi(\boldsymbol{x}_i)^T \Phi(\boldsymbol{x}_j) = \boldsymbol{x}_i^T \boldsymbol{x}_j + \frac{\delta_{ij}}{C}$$

where δ_{ij} is Kronecker's delta. In the kernel space the data points are again located on the edges of a regular simplex since

$$\begin{aligned} ||\Phi(\boldsymbol{x}_i)|| &= \sqrt{K(\boldsymbol{x}_i, \boldsymbol{x}_i)} = \sqrt{\boldsymbol{x}_i^T \boldsymbol{x}_i + \frac{1}{C}} = \sqrt{d + \frac{1}{C}}\,, \\ ||\Phi(\boldsymbol{x}_i) - \Phi(\boldsymbol{x}_j)|| &= \sqrt{\left(\Phi(\boldsymbol{x}_i) - \Phi(\boldsymbol{x}_j)\right)^T \left(\Phi(\boldsymbol{x}_i) - \Phi(\boldsymbol{x}_j)\right)} \\ &= \sqrt{K(\boldsymbol{x}_i, \boldsymbol{x}_i) - 2K(\boldsymbol{x}_i, \boldsymbol{x}_j) + K(\boldsymbol{x}_j, \boldsymbol{x}_j)} \\ &= \sqrt{\boldsymbol{x}_i^T \boldsymbol{x}_i + \boldsymbol{x}_j^T \boldsymbol{x}_j + \frac{2}{C}} = \sqrt{2\left(d + \frac{1}{C}\right)} \end{aligned}$$

and all samples are pairwise orthogonal:

$$\Phi(\boldsymbol{x}_i)^T \Phi(\boldsymbol{x}_j) = K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{x}_i^T \boldsymbol{x}_j = 0\ .$$

So, when training a 2-norm soft margin SVM we intrinsically deal with a hard-margin SVM of dimensionality $(d+\frac{1}{C})$. For $C \gg 1$ the softness term is negligible but for $C < 1$ dimensionality will increase dramatically. Thus, with the soft-margin approach we implicitly increase dimensionality so that the leave-one-out error rate will be close to 1 already for a smaller number of dimensions.

At first glance, this is counterintuitive since soft margin approaches increase the margin and therefore reduce the fat-shattering dimension so that overfitting is reduced and generalisation performance improves. But in case of high-dimensional low sample size data the asymmetries of leave-one-out cross-validation stronger affect generalisation performance than overfitting problems do.

For 1-norm soft margin approaches we expect the same behaviour but a closed mathematical formulation cannot be derived as simply as above.

4 Conclusions

Machine learning methods provide good results easily as long as large sample sizes in connection with comparatively low dimensionality are given. However, in practical applications such as the analysis of microarray data or other high-dimensional data mining problems, we often have a reverse situation: Extremely few points, for which it is not possible to increase the sample size significantly and often a high-dimensional feature space resulting from massively parallel data acquisition.

Increasing the dimensionality of artificial randomly distributed small sample size data can be shown to result in a surprising reorganisation of the data on the vertices of a regular simplex (see Theorem 1). Exactly the same is happening for non-random data, i.e. for true 2-class problems whenever the two classes fulfil the proximity criterion of Theorem 2. Thus, for $d \to \infty$, random and non-random data scenarios are not to distinguish anymore – by any metric-based measure.

We examined the practical impact of $d \to \infty$, i.e. how *large* infinity really is. Our empirical simulations suggest that there might be a sub-exponential relation between sample size and dimensionality. For an identically distributed equal-sized two-class dataset consisting of 20 samples in 10000 dimensions we already have an infinity-like behaviour. Here, we cannot learn anything from the data by metric-based methods at all as distances are becoming approximately equal. Apparently, infinite dimensionality is not that *large*.

In general, we do not know the true variances and true mean values of the class distributions, so these values have to be estimated from the data. Again, due to the counterintuitive properties of high-dimensional low sample size data, the data might degenerate in just the same way for $d \to \infty$. Thus, whenever acquiring degenerated data we can only speculate whether we have random or non-random data. Although we focused on scenarios that are not learnable by definition, the results are truly important in real-world scenarios. Avoiding an asymmetric resampling scheme such as leave-one-out cross-validation may not be possible due to the small sample size.

We showed that a soft-margin approach does not improve the generalisation performance of the SVM on high-dimensional low sample size data. We gave a proof for the first data scenario that introducing softness has the same effect as an increase of dimensionality. Thus, the soft-margin technique is making infinity even *smaller*.

Especially biometric problems may suffer from artifacts of high-dimensional low sample size data as we showed exemplarily for leave-one-out cross-validation with support vector classification. Typically, these problems are supposed to have a much lower intrinsic dimension than the observed dimension; thus, it should be possible to reduce dimensionality to avoid the mentioned effects. Future work should cover critical validation techniques of a widespread range of machine learning methods to further investigate counterintuitive artifacts and to find solutions to overcome these effects.

References

1. Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J., Caligiuri, M., Bloomfield, C., Lander, E.: Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. Science 286(5439), 531–537 (1999)
2. Lockhart, D.J., Winzeler, E.: Genomics, Gene Expression and DNA Arrays. Nature 405, 827–836 (2000)
3. Cristianini, N., Shawe-Taylor, J.: Support Vector Machines and other kernel-based learning methods. Cambridge University Press, Cambridge (2000)
4. Martinetz, T., Labusch, K., Schneegaß, D.: SoftDoubleMinOver: A Simple Procedure for Maximum Margin Classification. In: Duch, W., Kacprzyk, J., Oja, E., Zadrożny, S. (eds.) ICANN 2005. LNCS, vol. 3697, pp. 301–306. Springer, Heidelberg (2005)
5. Kohavi, R.: A study of cross-validation and bootstrap for accuracy estimation and model selection. In: IJCAI, pp. 1137–1145 (1995)
6. Hall, P., Marron, J.S., Neeman, A.: Geometric representation of high dimension, low sample size data. J. R. Statist. Soc. 67(3), 427–444 (2005)
7. Ahn, J., Marron, J.S., Muller, K.M., Chi, Y.Y.: The high-dimension, low-sample-size geometric representation holds under mild conditions. Biometrika 94(3), 760–766 (2007)
8. Bartlett, P., Shawe-Taylor, J.: Generalization Performance of Support Vector Machines and Other Pattern Classifiers. In: Advances in Kernel Methods: Support Vector Learning, pp. 43–54. MIT Press, Cambridge (1999)