Understanding Support Vector Machines with Polynomial Kernels

Rikard Vinge
Department of Electrical Engineering
Chalmers University of Technology
Gothenburg, Sweden
vinge@chalmers.se

Tomas McKelvey
Department of Electrical Engineering
Chalmers University of Technology
Gothenburg, Sweden
tomas.mckelvey@chalmers.se

***Abstract*—Interpreting models learned by a support vector machine (SVM) is often difficult, if not impossible, due to working in high-dimensional spaces. In this paper, we present an investigation into polynomial kernels for the SVM. We show that the models learned by these machines are constructed from terms related to the statistical moments of the support vectors. This allows us to deepen our understanding of the internal workings of these models and, for example, gauge the importance of combinations of features. We also discuss how the SVM with a quadratic kernel is related to the likelihood-ratio test for normally distributed populations.**

***Index Terms*—Interpretation, Support Vector Machine, Polynomial Kernel, Statistical Moments, Likelihood Ratio Test, Quadratic Discrimination**

I. Introduction

Support vector machines (SVMs) are attractive due to their many useful properties, including efficient training algorithms and proven performance on a multitude of different kinds of real-world inference problems. What they lack, however, is an ability to provide understandable details on the trained models. This lack of interpretability is common among many of the popular machine learning algorithms of today. In the hunt for better performance, algorithms that are flexible and generalize well are often preferred over algorithms whose inner mechanisms can be easily understood.

On the other end of the spectrum are simple, but often remarkably effective [1], learners such as Linear or Quadratic Discriminant Analysis (LDA and QDA, respectively), decision trees, naive Bayes classifiers, linear regression, and generalized additive methods [2]. Although they may not yield state-of-the-art performance in many applications, they offer the means to understanding the trained models and give explanations to how they arrived at a particular output.

Throughout the years, much work has been put into defining interpretability of machine learning models [3]–[5] and designing methods that provide interpretation to models from any learning algorithm [6]–[9]. Learner-specific interpretation models have been developed for many different kinds of machine learning methods, e.g. for artificial neural networks [10]–[13], random forests [14], and SVM [15].

In this paper, we investigate the structure of solutions produced by the SVM with polynomial kernels and show that the coefficients of the polynomial decision functions are related to the statistical moments of the support vectors. This allows for a deeper understanding of the SVM solution in terms of, for example, correlation between elements of the support vectors.

Specifically, we find a relationship between the model learned by SVM with quadratic kernels and QDA, with its basis in the likelihood ratio test. This relationship is not obvious from the way the two classifiers are constructed. In the case of the SVM, we directly learn a decision function

$$f(x) = \beta^T \phi(x) + \beta_0 \underset{\hat{y}=-1}{\overset{\hat{y}=+1}{\gtrless}} 0,$$

where $\hat{y}$ is the predicted class of sample x, from training data $(y_1, x_1), \ldots, (y_N, x_N)$, $y_i \in \{+1, -1\}$ and $x_i \in \mathcal{R}^p$. Quadratic kernels corresponds to transformations $\phi(x)$ with squared and cross combinations of the elements of x as well as the elements in x themselves and possibly a constant bias. For separable problems, the SVM is trained by solving the optimization function

$$\begin{aligned} \max_{\beta,\beta_0} \quad & M \\ \text{s.t.} \quad & \\ & y_i f(x_i) \geq M, \forall i = 1, \ldots, N \\ & ||\beta||_2 = 1 \end{aligned}$$

Here, M denotes the margin of the hyperplane $f(x)$, i.e. the distance from the hyperplane to the most adjacent training samples. At this point, the margins can be scaled by scaling β and for convenience the margin is set to $M = 1/||\beta||_2$ and the constraint on $||\beta||_2$ is removed. This optimization problem can be solved easily in its dual formulation [16] and produces solutions of the form

$$\beta = \sum_{i=1}^{N} \alpha_i y_i \phi(x_i)$$

with $\alpha_i > 0$ only for samples with $y_i f(x_i) = M$ and $\alpha_i = 0$ otherwise. This procedure of maximizing the margins for separable problems minimizes the structural risk [17], [18] and provides good generalization properties for the SVM. In the soft-margin version of the SVM, used for non-separable problems, the structural risk is not necessarily minimized.

978-9-0827-9703-9/19/$31.00

Instead, the bias and variance terms of the classifier are minimized by balancing the width of the margin with a penalization of erronous classifications.

In contrast to the SVM, which does not consider the distribution of the data, QDA assumes normally distributed data and uses the generalized likelihood ratio test to distinguish between them. For known distributions, the likelihood ratio test

$$L(x) = \frac{P(\hat{y}=+1|x)}{P(\hat{y}=-1|x)} \underset{\hat{y}=-1}{\overset{\hat{y}=+1}{\gtrless}} \gamma$$

has the highest detection rate for any probability of false alarm [19]. For two normally distributed classes, with means μ_{+1} and μ_{-1} and covariances Σ_{+1} and Σ_{-1} and the probability density function

$$P(x, \mu_{\pm 1}, \Sigma_{\pm 1}) = \frac{\exp\left(-\frac{1}{2}(x-\mu_{\pm 1})^T \Sigma_{\pm 1}^{-1}(x-\mu_{\pm 1})\right)}{\sqrt{(2\pi)^p |\Sigma_{\pm 1}|}},$$

the likelihood ratio is

$$L(x) = \frac{P(x, \mu_{+1}, \Sigma_{+1})}{P(x, \mu_{-1}, \Sigma_{-1})}.$$

For simplicity, the equivalent log likelihood ratio test is used for the normal distributions, in which the decision function becomes quadratic,

$$f(x) = x^T B x + b^T x + b_0$$

with

$$B = \frac{1}{2}\left(\Sigma_{-1}^{-1} - \Sigma_{+1}^{-1}\right) \tag{1}$$

$$b = \Sigma_{+1}^{-1}\mu_{+1} - \Sigma_{-1}^{-1}\mu_{-1}. \tag{2}$$

Distributions with unknown parameters are handled by the generalized likelihood ratio test. Assuming the data of the two classes are normal, we use the maximum likelihood estimators for the sample mean $\hat{\mu}$ and sample covariance $\hat{\Sigma}$, where

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i, \text{ and}$$

$$\hat{\Sigma} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \hat{\mu})(x_i - \hat{\mu})^T.$$

Replacing μ and Σ in (1) and (2) with $\hat{\mu}$ and $\hat{\Sigma}$, respectively, is the QDA machine learning algorithm [20], first applied in [21]. If the covariances are assumed equal, QDA reduces to LDA. In this case, the decision function is linear, similar to the one learned by linear SVM.

The remainder of this paper is organized as follows: in Section II we study the structure of SVMs with a simple quadratic kernel and in Section III we modify the kernel to the common polynomial kernel of degree two. The results are generalized to higher order kernels in Section IV. The paper is concluded in Section V.

II. Support vector machines with a simple quadratic kernel

Before looking at the most common quadratic kernel used today, we start with a simpler kernel stemming from the quadratic decision function we find in QDA,

$$f(x) = x^T B x + b^T x + b_0$$

where $x \in \mathbb{R}^p$ and the coefficients $B \in \mathbb{R}^{p\times p}$, $b \in \mathbb{R}^p$ and $b_0 \in \mathbb{R}$. Vectorizing and combining the terms gives the equivalent function

$$f(x) = \begin{bmatrix} \text{vec}(B) \\ b \end{bmatrix}^T \begin{bmatrix} x \otimes x \\ x \end{bmatrix} + b_0.$$

This is a linear function,

$$f(x) = \beta^T \phi_s(x) + b_0,$$

in the transformed, quadratic, space

$$\phi_s(x) = \begin{bmatrix} x \otimes x \\ x \end{bmatrix} = \tag{3}$$

$$= \left[x_1^2, \ldots, x_p^2, x_p x_{p-1}, \ldots\ldots x_1 x_2, x_p, \ldots, x_1\right]^T,$$

to which we can apply the SVM. The transformation $\phi_s(x)$ is the simplest quadratic transformation that contains all square-, cross- and linear combinations of the elements of x. The bias term b_0 is left out of the transformation to keep in line with the standard formulation of the SVM.

Solving the SVM in the transformed space, ϕ_s, yield solutions of the form

$$\beta = \begin{bmatrix} \text{vec}(B) \\ b \end{bmatrix} = \sum_{i=1}^{N} \alpha_i y_i \phi_s(x_i). \tag{4}$$

Converting the terms of (4) back to the quadratic and linear parts of the decision function, we find the coefficients B and b as

$$B = \sum_{i=1}^{N} \alpha_i y_i x_i x_i^T \tag{5}$$

$$b = \sum_{i=1}^{N} \alpha_i y_i x_i. \tag{6}$$

The linear coefficient is identical to that of the linear SVM. The quadratic coefficient has the equivalent form

$$B = \sum_{\{i:y_i=+1\}} \alpha_i x_i x_i^T - \sum_{\{i:y_i=-1\}} \alpha_i x_i x_i^T. \tag{7}$$

This is the difference between scaled sample correlation matrices of the support vectors from the two classes, where the scales are the support vector coefficients, here denoted support vector correlation matrices. Furthermore, the support vector correlation matrices are rank deficient if the number of support vectors are less than the dimension of B, in-line with the sparseness property of the SVM. The structure of the indefinite matrix B is similar to that of the corresponding term (1) in the solution to QDA. Three important differences between the SVM and QDA solutions are: the inversion of the covariance

matrices of the QDA; the inverted order of the subtraction of the SVM; and the fact that QDA uses the covariance while correlation matrices appear in the SVM solution. For zero-mean classes, the similarities are greater. We reiterate that QDA assumes the two classes to be normally distributed and requires at least $p+1$ samples per class in order to estimate the sample covariance matrices. Conversely, the SVM makes no assumptions on the distribution of the classes and requires only that support vectors of each class exists, which is guaranteed.

Collecting the terms of the decision function of QDA we have

$$\begin{aligned} f(x) =& x^T \hat{\Sigma}_{-1}^{-1} x - \hat{\mu}_{-1}^T \hat{\Sigma}_{-1}^{-1} x \\ & - x^T \hat{\Sigma}_{+1}^{-1} x + \hat{\mu}_{+1}^T \hat{\Sigma}_{+1}^{-1} x + b_0 \end{aligned}$$

and for SVM

$$\begin{aligned} f(x) =& x^T \left(\sum_{\{i:y_i=+1\}} \alpha_i x_i x_i^T \right) x + \left(\sum_{\{i:y_i=+1\}} \alpha_i x_i^T \right) x \\ & - x^T \left(\sum_{\{i:y_i=-1\}} \alpha_i x_i x_i^T \right) x - \left(\sum_{\{i:y_i=-1\}} \alpha_i x_i^T \right) x \\ & + b_0 \end{aligned}$$

Although QDA models the population distributions before distinguishing them and the SVM learns the decision function directly, the two functions are similar. One interpretation of this is that the SVM internally tries to sparsely model the population distributions as some distribution similar to the Gaussian, while simultaneously maximizing the margins.

If we look at the margin of the SVM we find

$$M = 1/\sqrt{||B||_F^2 + ||b||^2},$$

where $||.||_F$ denotes the Frobenius norm. Maximizing the margins equates to minimizing the sum of the norms of the two coefficients.

III. The common quadratic kernel

Quadratic decision functions for the SVM are most commonly achieved by the kernel function

$$K_q(x_i, x_j, r) = \left(x_i^T x_j + r\right)^2, \tag{8}$$

with the corresponding transformation

$$\begin{aligned} &\phi_q(x,r)^T = \\ &\left[x_1^2, \ldots, x_p^2, \sqrt{2}x_p x_{p-1}, \ldots \sqrt{2}x_1 x_2, \sqrt{2r}x_p, \ldots, \sqrt{2r}x_1, r\right]. \end{aligned} \tag{9}$$

Solutions to the SVM with this transformation is not as straightforward to interpret as with the simpler quadratic transformation, but reshuffling of the terms in (9) to the same order as in (3) and appending the bias r to the simple transformation allows us to write

$$\phi_q(x) = \begin{bmatrix} I_p & 0 & 0 & 0 \\ 0 & \sqrt{2}I_{p(p-1)} & 0 & 0 \\ 0 & 0 & \sqrt{2r}I_p & 0 \\ 0 & 0 & 0 & r \end{bmatrix} \begin{bmatrix} \phi_s(x) \\ 1 \end{bmatrix}.$$

Here, I_p and $I_{p(p-1)}$ are the $p \times p$ and $p \times (p-1)$ identity matrices, respectively. Appending the constant bias term, r, to transformation ϕ_s has no effect on the solution of the SVM, due to the constraint on the support vector coefficients that

$$\sum_{i=1}^{N} \alpha_i y_i = 0.$$

Converting the solution to the SVM with the common quadratic kernel back to the original space yields

$$B = \sum_{i=1}^{N} \alpha_i y_i \left(T \circ \left(x_i x_i^T\right)\right) \tag{10}$$

$$b = \sqrt{2r} \sum_{i=1}^{N} \alpha_i y_i x_i, \tag{11}$$

where

$$T_{ij} = \begin{cases} 1, & i = j \\ \sqrt{2}, & i \neq j \end{cases} \tag{12}$$

and $\circ$ denotes the Hadamard product. The quadratic term now consists of the same difference between support vector correlation matrices as (5), but scaled element-wise by T, more clearly seen in the equivalent formulation

$$B = T \circ \sum_{i=1}^{N} \alpha_i y_i x_i x_i^T.$$

Expressing the margin in terms of the B and b found from transformation (3) we find that the bias term r from (8) allows for control over the relative importance of the quadratic and linear coefficients in the decision functions,

$$M = 1/\sqrt{||B||_F^2 + ||b||^2} = 1/\sqrt{||T \circ B_s||_F^2 + 2r||b_s||^2}.$$

Here, B_s and b_s are the solutions (5) and (6). Thus, SVMs solved with the common quadratic kernels relates to QDA in a similar fashion as the simple quadratic transformation. The element-wise scaling by T assigns higher importance on the cross-terms of the support vector correlation matrices, compared to the diagonal terms. The general conclusions on the relationship between the SVM and QDA holds for the common quadratic kernel as for the simpler, but the similarity between the quadratic terms in the two classification methods is reduced. The importance of the bias r becomes clear from the margin, where it acts as a weight between the linear and quadratic term.

IV. Higher order polynomial kernels

Higher-order polynomial kernels can be handled similarly to the quadratic. SVMs with polynomial kernels of order three learns decision functions of the form

$$f(x) = \sum_{k=1}^{p} x^k x^T \mathcal{B}^k x + x^T B x + b^T x + b_0,$$

where $\mathcal{B}^k$ is the kth matrix of the third-order tensor $\mathcal{B}$ and x^k the kth element of x. For the cubic transformation

$$\phi_c(x) = \begin{bmatrix} x \otimes x \otimes x \\ x \otimes x \\ x \end{bmatrix}$$

we find the solution

$$\mathcal{B}^k = \sum_{i=1}^{N} \alpha_i y_i x_i^k x_i x_i^T, \quad k = 1, \ldots, p \tag{13}$$

$$B = \sum_{i=1}^{N} \alpha_i y_i x_i x_i^T \tag{14}$$

$$b = \sum_{i=1}^{N} \alpha_i y_i x_i. \tag{15}$$

The matrix $\mathcal{B}^k$ constitutes the difference

$$\mathcal{B}^k = \sum_{\{i:y_i=+1\}} \alpha_i x_i^k x_i x_i^T - \sum_{\{i:y_i=-1\}} \alpha_i x_i^k x_i x_i^T$$

where the tensors are related to the, non-standardized, coskewness of the support vectors. The corresponding solution with the common third-order polynomial kernel

$$K_c(x_i, x_j, r) = \left(x_i^T x_j + r\right)^3$$

is

$$\mathcal{B}^k = \sum_{i=1}^{N} \alpha_i y_i \left(\mathcal{T}^k \circ x_i^k x_i x_i^T\right), \quad k = 1, \ldots, p$$

$$B = \sqrt{3r} \sum_{i=1}^{N} \alpha_i y_i \left(T \circ \left(x_i x_i^T\right)\right)$$

$$b = \sqrt{3} r \sum_{i=1}^{N} \alpha_i y_i x_i.$$

The elements of $\mathcal{T}^k$ are 1 for the cubic terms and $\sqrt{3}$ for all cross terms. The elements of T is the same as in (12). The margin for the cubic kernel is

$$M = \sqrt{||\text{vec}\left(\mathcal{T} \circ \mathcal{B}_s\right)||_2^2 + 3r||\text{vec}\left(T \circ B_s\right)||_2^2 + 3r^2||b_s||_2^2},$$

where $\mathcal{B}_s$, B_s and b_s are the solutions (13)-(15) for the simple cubic transformation. The element order of the tensor vectorization operator is ill-defined but is unimportant in this case. We also find that the weighting imposed by r on the terms of the margin is given by the binomial coefficients.

Extension to fourth order and higher polynomials is straightforward and we find that the degree-four component of the decision function is related to the, non-standardized, cokurtosis of the support vectors, in the same manner as the third-order component is related to coskewness and the second-order to correlation.

In general, the d-degree component of the decision function is related to the dth moment of each class in the training set. For an n-degree polynomial kernel, the weight of the dth term of the $n-1$ terms of the margin is

$$\binom{n}{d} r^d.$$

For high-degree polynomials, even small deviations from $r = 1$ will affect the lower-order terms greatly.

V. Conclusion

We have shown that models trained with an SVM using a polynomial kernel can be expressed in compact form as polynomials with tensor coefficients computed only from the support vectors. These polynomial coefficients are related to scaled sample moments of the support vectors, where the scales are the support vector coefficients multiplied by constants determined by the polynomial degree. In contrast to the statistical moments, the coefficients of the polynomial SVM model are neither centered to zero mean nor standardized to unit variance and, thus, are not scale invariant. For the quadratic kernel, the learned model consists of one part identical to the standard linear SVM and one extra term computed from scaled correlation matrices of the support vectors from the two classes. The structure of the learned model bears resemblance to models learned by QDA, with the stark difference that QDA requires full-rank sample covariance matrices while the SVM requires only a sparse collection of support vectors.

This compact form of the polynomial SVM means that interpretation of the models is possible, as dominating terms in any of the tensor coefficients are directly related to combinations between the features in the original space. For the quadratic case, we can look for strong correlations between elements of the support vectors, identically to how we search for strong terms in the inverted covariance matrices of QDA in order to explain its results.

Furthermore, we have shown that the margin of an SVM with a polynomial kernel SVM can be expressed in terms of the coefficients of the learned polynomial decision function, weighted by powers of the bias r of the polynomial kernel and the binomial coefficients.

The analysis used in this report can be applied to any kernel that can be expressed in terms of finite series of polynomials, e.g. a truncated polynomial expansion of the radial basis function kernel, and provide a means to better understand these SVM models. Also, the compact form of the polynomial SVM can be used to store the SVM classifier without storing the support vectors and may have uses where it is less computationally or memory intensive compared to the non-compact form.

References

[1] D. Michie, D. J. Spiegelhalter, and C. C. Taylor, *Machine Learning, Neural and Statistical Classification*. Ellis Horwood, 1994.

[2] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad, "Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day readmission," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015, pp. 1721–1730. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2783258.2788613

[3] W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu, "Interpretable machine learning: definitions, methods, and applications," *arXiv:1901.04592v1 [stat.ML]*, pp. 1–11, 2019. [Online]. Available: http://arxiv.org/abs/1901.04592

[4] F. Doshi-Velez and B. Kim, "Towards A Rigorous Science of Interpretable Machine Learning," *arXiv:1702.08608v2 [stat.ML]*, 2017. [Online]. Available: http://arxiv.org/abs/1702.08608

[5] Z. C. Lipton, "The Mythos of Model Interpretability," in *Proceedings of the 2016 ICML Workshop on Human Interpretability of Machine Learning*, 2016, pp. 96–100. [Online]. Available: http://arxiv.org/abs/1606.03490

[6] J. D. Olden and D. A. Jackson, "Illuminating the "black box": A randomization approach for understanding variable contributions in artificial neural networks," *Ecological Modelling*, vol. 154, no. 1-2, pp. 135–150, 2002.

[7] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why Should I Trust You?" Explaining the Predictions of any Classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144.

[8] J. Krause, A. Perer, and E. Bertini, "Using Visual Analytics to Interpret Predictive Machine Learning Models," in *Proceedings of the 2016 ICML Workshop on Human Interpretability of Machine Learning*, 2016, pp. 106–110. [Online]. Available: http://arxiv.org/abs/1606.05685

[9] P. Ponte and R. G. Melko, "Kernel methods for interpretable machine learning of order parameters," *Physical Review B*, vol. 96, no. 20, p. 205146, 2017.

[10] D. G. Garson, "Interpreting neural-network connection strengths," *AI Expert*, vol. 6, no. 4, pp. 46–51, 1991.

[11] O. Intrator and N. Intrator, "Interpreting Neural-Network Results: a simulation study," *Computational Statistics and Data Analysis*, vol. 37, pp. 373–393, 2001.

[12] N. Frosst and G. Hinton, "Distilling a Neural Network Into a Soft Decision Tree," *arXiv:1711.09784v1 [cs.LG]*, 2017.

[13] M. W. Beck, "NeuralNetTools: Visualization and Analysis Tools for Neural Networks," *Journal of Statistical Software*, vol. 85, no. 11, pp. 1–20, 2018. [Online]. Available: http://www.jstatsoft.org/v85/i11/

[14] K. J. Archer and R. V. Kimes, "Empirical characterization of random forest variable importance measures," *Computational Statistics and Data Analysis*, vol. 52, no. 4, pp. 2249–2260, 2008.

[15] T. Maszczyk and D. Wlodzislaw, "Support Vector Machines for Visualization and Dimensionality Reduction," in *Proceedings of the 18th International Conference on Artificial Neural Networks*, 2008, pp. 346–356. [Online]. Available: http://link.springer.com/10.1007/978-3-540-87536-9

[16] C. Cortes and V. N. Vapnik, "Support-Vector Networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[17] V. Vapnik, "Principles of risk minimization for learning theory," in *Proceedings of the 4th International Conference on Neural Information Processing Systems*, 1991, pp. 831–838. [Online]. Available: http://papers.nips.cc/paper/506-principles-of-risk-minimization-for-learning-theory

[18] I. M. Guyon, V. N. Vapnik, B. E. Boser, L. Bottou, and S. A. Solla, "Structural risk minimization for character recognition," in *Proceedings of Advances in Neural Information Processing Systems 4*, 1992, pp. 471–479.

[19] J. Neyman and E. S. Pearson, "On the Problem of the Most Efficient Tests of Statistical Hypotheses," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 231, no. 694-706, pp. 289–337, 1933. [Online]. Available: http://rsta.royalsocietypublishing.org/cgi/doi/10.1098/rsta.1933.0009

[20] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. Springer-Verlag, New York, 2009. [Online]. Available: http://www-stat.stanford.edu/ tibs/ElemStatLearn/

[21] R. A. Fisher, "The Use of Multiple Measurements in Taxonomic Problems," *Annals of Eugenics*, vol. 7, no. 2, pp. 179–188, 1936. [Online]. Available: http://doi.wiley.com/10.1111/j.1469-1809.1936.tb02137.x