Performance analysis for L_2 kernel classification

Authored by:

Clayton Scott
Jooseuk Kim

Abstract

We provide statistical performance guarantees for a recently introduced kernel classifier that optimizes the L_2 or integrated squared error (ISE) of a difference of densities. The classifier is similar to a support vector machine (SVM) in that it is the solution of a quadratic program and yields a sparse classifier. Unlike SVMs, however, the L_2 kernel classifier does not involve a regularization parameter. We prove a distribution free concentration inequality for a cross-validation based estimate of the ISE, and apply this result to deduce an oracle inequality and consistency of the classifier on the sense of both ISE and probability of error. Our results can also be specialized to give performance guarantees for an existing method of L_2 kernel density estimation.

1 Paper Body

In the binary classification problem we are given realizations (x1 , y1), . . . , (xn , yn) of a jointly distributed pair (X, Y), where X ? Rd is a pattern and Y ? {?1, +1} is a class label. The goal of classification is to build a classifier, i.e., a function taking X as input and outputting a label, such that some measure of performance is optimized. Kernel classifiers [1] are an important family of classifiers that have drawn much recent attention for their ability to represent nonlinear decision boundaries and to scale well with increasing dimension d. A kernel classifier (without offset) has the form (n) X g(x) = sign ?i yi k(x, xi) , i=1

where ?i are parameters and k is a kernel function. For example, support vector machines (SVMs) without offset have this form [2], as does the standard kernel density estimate (KDE) plug-in rule. Recently Kim and Scott [3] introduced an L2 or integrated squared error (ISE) criterion to design the coefficients ?i of a kernel classifier with Gaussian kernel. Their L2 classifier performs comparably to existing kernel methods while possesing a number of desirable properties. Like the SVM, L2 kernel classifiers are the solutions of convex quadratic programs that can be solved efficiently using standard decomposition algorithms. In addition, the classifiers are sparse, meaning most of the coefficients ?i = 0, which has advantages for representation and evaluation efficiency. Unlike the SVM, however, there are no free parameters to be set by the user except the kernel bandwidth parameter. In this paper we develop statistical performance guarantees for the L2 kernel classifier introduced in [3]. The linchpin of our analysis is a new concentration inequality bounding the deviation of a cross-validation based ISE estimate from the true ISE. This bound is then applied to prove an oracle inequality and consistency in both ISE and probability of error. In addition, as a special case of ?

Both authors supported in part by NSF Grant CCF-0830490

1

our analysis, we are able to deduce performance guarantees for the method of L2 kernel density estimation described in [4, 5]. The ISE criterion has a long history in the literature on bandwidth selection for kernel density estimation [6] and more recently in parametric estimation [7]. The use of ISE for optimizing the weights of a KDE via quadratic programming was first described in [4] and later rediscovered in [5]. In [8], an ‘1 penalized ISE criterion was used to aggregate a finite number of pre-determined densities. Linear and convex aggregation of densities, based on an L2 criterion, are studied in [9], where the densities are based on a finite dictionary or an independent sample. In contrast, our proposed method allows data-adaptive kernels, and does not require and independent (holdout) sample. In classification, some connections relating SVMs and ISE are made in [10], although no new algorithms are proposed. Finally, the ?difference of densities? perspective has been applied to classification in other settings by [11], [12], and [13]. In [11] and [13], a difference of densities are used to find smoothing parameters or kernel bandwidths. In [12], conditional densities are chosen among a parameterized set of densities to maximize the average (bounded) density differences. Section 2 reviews the L2 kernel classifier, and presents a slight modification needed for our analysis. Our results are presented in Section 3. Conclusions are offered in the final section, and proofs are gathered in an appendix.

2

L2 Kernel Classification

We review the previous work of Kim & Scott [3] and introduce an important modification. For convenience, we relabel Y so that it belongs to {1, ??} and denote I+ = {i — Yi = +1} and I? = {i — Yi = ??}. Let f? (x) and f+ (x) denote the class-conditional densities of the pattern given the label. From decision theory, the optimal classifier has the form g ? (x) = sign {f+ (x) ? ?f? (x)} ,

(1)

where ? incorporates prior class probabilities and class-conditional error costs (in the Bayesian setting) or a desired tradeoff between false positives and false negatives [14]. Denote the ?difference of densities? d? (x) := f+ (x) ? ?f? (x). The class-conditional densities are modelled using the Gaussian kernel as X X fb+ (x; ?) = ?i k? (x, Xi) , fb? (x; ?) = ?i k? (x, Xi) i?I+

i?I?

with constraints ? = (?1 , . . . , ?n) ? A where ? ? X X ?i = ?i = 1, A= ?— ? i?I+

i?I?

?i ? 0

? ? ?i . ?

The Gaussian kernel is defined as

? ? ? ??d/2 kx ? Xi k2 . k? (x, Xi) = 2?? 2 exp ? 2? 2

The ISE associated with ? is

Z ? ?2 ISE (?) = kdb? (x; ?) ? d? (x) k2L2 = db? (x; ?) ? d? (x) dx Z Z Z 2 b b = d? (x; ?) dx ? 2 d? (x; ?) d? (x) dx + d2? (x) dx.

Since we do not know the true d? (x), we need to estimate the second term in the above equation Z H (?) , db? (x; ?) d? (x) dx (2) by Hn (?) which will be explained in detail in Section 2.1. Then, the empirical ISE is Z Z d (?) = db2? (x; ?) dx ? 2Hn (?) + d2? (x) dx. ISE 2

(3)

b is defined as Now, ?

d (?) b = arg min ISE ? ??A

and the final classifier will be

2.1

(+1, g (x) = ??,

(4)

b ?0 db? (x; ?) b b ¡ 0. d? (x; ?)

Estimation of H (?)

In this section, we propose a method of estimating H (?) in (2). The basic idea is to view H (?) as an expectation and estimate it using a sample average. In [3], the resubstitution estimator for H (?) was used. However, since this estimator is biased, we use a leave-one-out cross-validation (LOOCV) estimator, which is unbiased and facilitates our theoretical analysis. Note that the difference of densities can be expressed as n X db? (x; ?) = fb+ (x) ? ? fb? (x) = ?i Yi k? (x, Xi) . i=1

Then,

Z H (?) = =

Z db? (x; ?) d? (x) dx =

Z X n

Z db? (x; ?) f+ (x) dx ? ?

?i Yi k? (x, Xi) f+ (x) dx ? ?

Z X n

i=1

=

n X

db? (x; ?) f? (x) dx

?i Yi k? (x, Xi) f? (x) dx

i=1

?i Yi h (Xi)

i=1

where

Z h (Xi) ,

Z k? (x, Xi) f+ (x) dx ? ?

k? (x, Xi) f? (x) dx.

(5)

We estimate each h (Xi) in (5) for i = 1, . . . , n using leave-one-out cross-validation ? X ? X 1 ? ? k? (Xj , Xi) ? k? (Xj , Xi) , i ? I+ ? ? N+ ? 1 N? j?I? j?I+ ,j6=i b hi , X ? 1 X ? ? k (X , X) ? k? (Xj , Xi) , i ? I? ? j i ? ? N+ N? ? 1 j?I+ j?I? ,j6=i Pn hi . where N+ = —I+ — , N? = —I? —. Then, the estimate of H (?) is Hn (?) = i=1 ?i Yi b 2.2

Optimization

The optimization problem (4) can be formulated as a quadratic program. The first term in (3) is !2 Z Z ?X n 2 ?i Yi k? (x, Xi) dx db? (x; ?) dx = i=1

=

n X n X

Z

?i ?j Yi Yj

k? (x, Xi) k? (x, Xj) dx =

i=1 j=1

n X n X

?i ?j Yi Yj k?2? (Xi , Xj)

i=1 j=1

by the convolution theorem for Gaussian kernels [15]. As we have seen in Section 2.1, the second Pn term Hn (?) in (3) is linear in ? and can be expressed as i=1 ?i ci where ci = Yi b hi . Finally, since the third term does not depend on ?, the optimization problem (4) becomes the following quadratic program (QP) n n n X 1 XX b = arg min ?i ?j Yi Yj k?2? (Xi , Xj) ? ci ?i . (6) ? 2 i=1 j=1 ??A i=1 The QP (6) is similar to the dual QP of the 2-norm SVM with hinge loss [2] and can be solved by a variant of the Sequential Minimal Optimization (SMO) algorithm [3]. 3

3

Statistical performance analysis

In this section, we give theoretical performance analysis on our proposed method. We assume that {Xi }i?I+ and {Xi }i?I? are i.i.d samples from f+ (x) and f? (x), respectively, and treat N+ and N? as deterministic variables n+ and n? such that n+ ? ? and n? ? ? as n ? ?. 3.1

Concentration inequality for Hn (?)

Lemma 1. Conditioned on Xi , b hi is an unbiased estimator of h (Xi), i.e, h i E b hi — Xi = h (Xi) . Furthermore, for any ? ¿ 0, ? ? ? ? 2 2 P sup —Hn (?) ? H (?)— ¿ ? ? 2n e?c(n+ ?1)? + e?c(n? ?1)? ??A where c = 2

?? ?2d 4 2?? / (1 + ?) .

Lemma 1 implies that Hn (?) ? H (?) almost surely for all ? ? A simultaneously, provided that ?, n+ , and n? evolve as functions of n such that n+ ? 2d / ln n ? ? and n? ? 2d / ln n ? ?. 3.2

Oracle Inequality

Next, we establish on oracle inequality, which relates the performance of our estimator to that of the best possible kernel classifier. ? ? 2 2 Theorem 1. Let

? ¿ 0 and set ? = ? (?) = 2n e?c(n+ ?1)? + e?c(n? ?1)? where c = ?2d ?? 4 / (1 + ?) . Then, with probability at least 1 ? ? 2 2?? b ? inf ISE (?) + 4?. ISE (?) ??A Proof. From Lemma 1, with probability at least 1 ? ? ? ? ? d (?)?? ? 2?, ?ISE (?) ? ISE

?? ? A

d (?) = 2 (Hn (?) ? H (?)). Then, with probability at least 1 ? ?, by using the fact ISE (?) ? ISE for all ? ? A, we have d (?) d (?) + 2? ? ISE (?) + 4? b ? ISE b + 2? ? ISE ISE (?) b This proves the theorem. where the second inequality holds from the definition of ?. 3.3

ISE consistency

b converges to zero in probability. Next, we have a theorem stating that ISE (?) Theorem 2. Suppose that for f = f+ and f? , the Hessian Hf (x) exists and each entry of Hf (x) is piecewise continuous and square integrable. If ?, n+ , and n? evolve as functions of n such that b ? 0 in probability as n ? ? ? ? 0, n+ ? 2d / ln n ? ?, and n? ? 2d / ln n ? ?, then ISE (?) This result intuitively follows from the oracle inequality since the standard Parzen window density estimate is consistent and uniform weights are among the simplex A. The rigorous proof is omitted due to space limitations. 4

3.4

Bayes Error Consistency

In classification, we are ultimately interested in minimizing the probability of error. Let us now n assume {Xi }i=1 is an i.i.d sample from f (x) = pf+ (x) + (1 ? p)f? (x), where 0 ¡ p ¡ 1 is the prior probability of the positive class. The consistency with respect to the probability of error could be easily shown if we set ? to ? ? = 1?p p and apply Theorem 3 in [17]. However, since p is ? unknown, we must estimate ? . Note that N+ and N? are binomial random variables, and we may ? estimate ? ? as ? = N N+ . The next theorem says the L2 kernel classifier is consistent with respect to the probability of error. Theorem 3. Suppose that the assumptions in Theorem 2 are satisfied. In addition, suppose that f? ? L2 (R), i.e. kf? kL2 ¡ ?. Let ? = N? /N+ be an estimate of ? ? = 1?p p . If ? evolves as 2d a function of n such that ? ? 0 and n? / ln n ? ? as n ? ?, then the L2 kernel classifier is consistent. In other words, given training data Dn = ((X1 , Y1) , . . . , (Xn , Yn)), the classification error n n o o b 6= Y — Dn Ln = P sgn db? (X; ?) converges to the Bayes error L? in probability as n ? ?. The proof is given in Appendix A.2. 3.5

Application to density estimation

By setting ? = 0, our goal becomes estimating f+ and we recover the L2 kernel density estimate of [4, 5] using leave-one-out cross-validation. Given an i.i.d sample X1 , . . . , Xn from f (x), the L2 kernel density estimate of f (x) is defined as b = fb(x; ?)

n X

? bi k? (x, Xi)

i=1

with ? bi ?s optimized such that b = arg ? min P ?i =1 ?i ?0

n n 1 XX

2

?i ?j k?2? (Xi , Xj) ?
i=1 j=1
n X i=1
?
? ?i ?
1 n?1
X
k? (Xi , Xj)? .
j6=i

Our concentration inequality, oracle inequality, and L2 consistency result immediately extend to provide the same performance guarantees for this method. For example, we state the following corollary. Corollary 1. Suppose that the Hessian Hf (x) of a density function f (x) exists and each entry of Hf (x) is piecewise continuous and square integrable. If ? ? 0 and n? 2d / ln n ? ? as n ? ?, then Z ? ?2 b ? f (x) dx ? 0 fb(x; ?) in probability.

4

Conclusion

Through the development of a novel concentration inequality, we have established statistical performance guarantees on a recently introduced L2 kernel classifier. We view the relatively clean analysis of this classifier as an attractive feature relative to other kernel methods. In future work, we hope to invoke the full power of the oracle inequality to obtain adaptive rates of convergence, and consistency for ? not necessarily tending to zero. 5

A A.1

Appendix Proof of Lemma 1

Note that for any given i, (k? (Xj , Xi))j6=i are independent and bounded by M = 1/ For random vectors Z ? f+ (x) and W ? f? (x), h (Xi) in (5) can be expressed as

?? ?d 2?? .

h (Xi) = E [k? (Z, Xi) — Xi] ? ?E [k? (W, Xi) — Xi] . Since Xi ? f+ (x) for i ? I+ and Xi ? f? (x) for i ? I? , it can be easily shown that h i E b hi — Xi = h (Xi) . For i ? I+ , ? ? ? ? ? ? ?b ? ? P hi ? h (Xi) ¿ ? ? Xi = x ? ? ? ?? X ? ? 1 ? ?? ? ? k? (Xj , Xi) ? E [k? (Z, Xi) — Xi]? ¿ Xi = x ? P ? n+ ? 1 1+? ? j?I+ ,j6=i ? ? ?? ? ? ? ? X ?? ?? ? ? k? (Xj , Xi) ? ?E [k? (W, Xi) — Xi]? ¿ + P ? Xi = x n? 1+? ? j?I?

By Hoeffding?s inequality [16], the first term in (7) is ? ? ?? X ? ? ? (n+ ? 1) ? ? ? ? ? P ? k? (Xj , Xi) ? (n+ ? 1)E [k? (Z, Xi) — Xi]? ¿ Xi = x 1+? ? j?I+ ,j6=i ? ?? X ? X ?? ? ? ? (n+ ? 1) ? ? ? ? ? = P ? k? (Xj , Xi) — Xi ? ¿ k? (Xj , Xi) ? E Xi = x 1+? ? j?I+ ,j6=i j?I+ ,j6=i ? ?? X ?? ? X ? ? ? (n+ ? 1) ? ? ? Xi = x = P ?? k? (Xj , Xi) — Xi ?? ¿ k? (Xj , Xi) ? E 1+? ? j?I+ ,j6=i
j?I+ ,j6=i
?
2e?2(n+ ?1)?
2
2

/(1+?) M

2

.

The second term in (7) is ? ? ?? X ? ? ? n? ? ?? P ?? X = x k? (Xj , Xi) ? n? E [k? (W, Xi) — Xi]?? ¿ i 1+? ? j?I? ? ? ?? X ?X ?? ? ? n? ? ?? ? ? Xi = x = P ? k? (Xj , Xi) ? E k? (Xj , Xi) — Xi ? ¿ 1+? ? j?I?

? 2e

j?I?

?2n? ?2 /(1+?)2 M 2

? 2e

?2(n? ?1)?2 /(1+?)2 M 2

.

Therefore, ? ? ?? ?? ? ? n? o ? ?b ? ?b ? ? P ?hi ? h (Xi)? ? ? = E P ?hi ? h (Xi)? ? ? ? Xi = X ? 2e?2(n+ ?1)?

2

/(1+?)2 M 2

+ 2e?2(n? ?1)?

2

/(1+?)2 M 2

.

In a similar way, it can be shown that for i ? I? , ? n? o 2 2 2 2 2 2 2 ? ? P ?b hi ? h (Xi)? ¿ ? ? 2e?2(n+ ?1)? /(1+?) M + 2e?2(n? ?1)? /(1+?) M .

6

(7)

Then, ? n ? ( ) ? ? ?X ? ?? ? ? P sup —Hn (?) ? H (?)— ¿ ? = P sup ? ?i Yi b hi ? h (Xi) ? ¿ ? ? ??A ??A ? i=1 ( ) n ? ? X ? ? ? P sup ?i —Yi — ?b hi ? h (Xi)? ¿ ? ??A i=1 ? ? n n ? ? X ? ? X ? ? ? ? = P sup ?i ?b hi ? h (Xi)? + ?i ? ?b hi ? h (Xi)? ¿ ? ??A i?I+ i?I? ? ? ? ? n n ? ? ? ? X ? X ? ?? ?? ? ?b ? b ? P sup B + P sup ? ? h ? h (X) ?i ?hi ? h (Xi)? ¿ ? ?¿ i i i ? 1+? 1+? ??A i?I+ ??A i?I? ? ? ? ? ? ? ? ? ? ? ? ? ?? ? ?? ? ? ? ? hi ? h (Xi)? ¿ hi ? h (Xi)? ¿ = P max ?b B + P max ?b B ? i?I+ i?I? 1+? 1+? ? ? [?? ?? ? ? [?? ?? ? ? ? ? ? ? ? ? ?b ? ?b ? ?B + P ?B ¿ = P h ? h (X) ? ?hi ? h (Xi)? ¿ ? i i ? 1+? 1+? ? i?I+ i?I? ? ? X ?? ? ? ? ? X ??? ? ?? ? ?? ? ? ?b b P ?hi ? h (Xi)? ¿ P ?hi ? h (Xi)? ¿ ? B + B 1+? ? 1+? ? i?I+ i?I? ? ? 2 4 2 2 4 2 ? n+ 2e?2(n+ ?1)? /(1+?) M + 2e?2(n? ?1)? /(1+?) M ? ? 2 4 2 2 4 2 + n? 2e?2(n+ ?1)? /(1+?) M + 2e?2(n? ?1)? /(1+?) M ? ? 2 4 2 2 4 2 = n 2e?2(n+ ?1)? /(1+?) M + 2e?2(n? ?1)? /(1+?) M . A.2

? ? ? ?B ?

Proof of Theorem 3

From Theorem 3 in [17], it suffices to show that Z ? ?2 b ? d? ? (x) dx ? 0 db? (x; ?) in probability. Since from the triangle inequality b ? d? ? (x) kL2 = kdb? (x; ?) b ? d? (x) + (? ? ? ? ?) f? (x) kL2 kdb? (x; ?) b ? d? (x) kL2 + k (? ? ? ?) f? (x) kL2 ? kdb? (x; ?) p b + —? ? ? ? — ? kf? (x) kL2 , = ISE (?) b and ? converges in probability to 0 and ? ? , respectively. The conwe need to show that ISE (?) ? vergence of ? to ? can be easily shown from the strong law of large numbers. b by treating N+ , N?

and ? In the previous analyses, we have shown the convergence of ISE (?) as deterministic n variables but now we turn to theocase where these variables are random. Define an n(1?p) , ? ? 2? ? . For any ? ¿ 0, event D = N+ ? np 2 , N? ? 2 ? ? ? ? b ¿ ?} ? P Dc + P ISE (?) b ¿ ?, D . P {ISE (?) The first term converges to 0 from the strong law of large numbers. Let define a set S = ? ? ? n(1?p) n? ? , n+ ? 2? ? . Then, (n+ , n?) n+ ? np 2 , n? ? 2 ? ? b ¿ ?, D P ISE (?) X ? ? ? b ¿ ?, D ? N+ = n+ , N? = n? ? P {N+ = n+ , N? = n? } = P ISE (?) X ? ? ? b ¿ ? ? N+ = n+ , N? = n? ? P {N+ = n+ , N? = n? } = P ISE (?) (n+ ,n?)?S

?

max

(n+ ,n?)?S

? ? ? b ¿ ? ? N + = n+ , N ? = n? . P ISE (?) 7

(8)

Provided that ? ? 0 and n? 2d / ln n ? ?, any pair (n+ , n?) ? S satisfies ? ? 0, n+ ? 2d / ln n ? ?, and n? ? 2d / ln n ? ? as n ? ? and thus the term in (8) converges to 0 from Theorem 2. This proves the theorem.

2 References

[1] B. Sch?olkopf and A. J. Smola, Learning with Kernels, MIT Press, Cambridge, MA, 2002. [2] C. Cortes and V. Vapnik, ?Support-vector networks,? Machine Learning, vol. 20, no. 3, pp. 273?297, 1995. [3] J. Kim and C. Scott, ?Kernel classification via integrated squared error,? IEEE Workshop on Statistical Signal Processing, August 2007. [4] D. Kim, Least Squares Mixture Decomposition Estimation, unpublished doctoral dissertation, Dept. of Statistics, Virginia Polytechnic Inst. and State Univ., 1995. [5] Mark Girolami and Chao He, ?Probability density estimation from optimally condensed data samples,? IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 25, no. 10, pp. 1253?1264, OCT 2003. [6] B.A. Turlach, ?Bandwidth selection in kernel density estimation: A review,? Technical Report 9317, C.O.R.E. and Institut de Statistique, Universit?e Catholique de Louvain, 1993. [7] David W.Scott, ?Parametric statistical modeling by minimum integrated square error,? Technometrics 43, pp. 274?285, 2001. [8] A.B. Tsybakov F. Bunea and M.H. Wegkamp, ?Sparse density estimation with l1 penalties,? Proceedings of 20th Annual Conference on Learning Theory, COLT 2007, Lecture Notes in Artificial Intelligence, v4539, pp. 530? 543, 2007. [9] Ph. Rigollet and A.B. Tsybakov, ?Linear and convex aggregation of density estimators,? https:// hal.ccsd.cnrs.fr/ccsd-00068216, 2004. [10] Robert Jenssen, Deniz Erdogmus, Jose C.Principe, and Torbj?rn Eltoft, ?Towards a unification of information theoretic learning and kernel method,? in Proc. IEEE Workshop on Machine Learning for Signal Processing (MLSP2004), Sao Luis, Brazil. [11] Peter Hall and Matthew P.Wand, ?On nonparametric discrimination using density differeces,? Biometrika, vol. 75, no. 3, pp. 541?547, Sept 1988. [12] P. Meinicke, T. Twellmann, and H. Ritter, ?Discriminative densities from maximum contrast

estimation,? in Advances in Neural Information Proceeding Systems 15, Vancouver, Canada, 2002, pp. 985?992. [13] M. Di Marzio and C.C. Taylor, ?Kernel density classification and boosting: an l2 analysis,? Statistics and Computing, vol. 15, pp. 113?123(11), April 2005. [14] E. Lehmann, Testing statistical hypotheses, Wiley, New York, 1986. [15] M.P. Wand and M.C. Jones, Kernel Smoothing, Chapman & Hall, 1995. [16] L. Devroye and G. Lugosi, ?Combinatorial methods in density estimation,? 2001. [17] Charles T. Wolverton and Terry J. Wagner, ?Asymptotically optimal discriminant fucntions for pattern classification,? IEEE Trans. Info. Theory, vol. 15, no. 2, pp. 258?265, Mar 1969.

8