E0 370 Statistical Learning Theory **Lecture 9 (Sep 13, 2011)**

Model Selection

Lecturer: Shivani Agarwal *Scribe: Shivani Agarwal*

1 Introduction

In the last few lectures, we have seen several techniques for bounding the generalization error of a function learned from a fixed function class (of limited capacity, or by a learning algorithm with good stability properties). Often, however, there is a choice of several possible function classes that can be used by an algorithm, generally with different capacities (think of choosing the number of nodes in a neural network, the height of a decision tree, or the degree of a polynomial kernel). Indeed, one can also use different algorithms to learn functions from different function classes. In such a situation, how should one choose the 'right' function class or model for the data? This is the classical problem of *model selection*.

There are several approaches that have been proposed for this problem in many different contexts and many different communities; we will not survey these here, but will mention a few commonly used approaches. Our focus will be on extending the techniques we have seen so far to bound the generalization error of a function learned not from a fixed function class, but from a function class picked by a model selection algorithm after seeing the data.[1] We will also apply similar ideas to obtain a margin-based generalization error bound that can be applied using a margin parameter value chosen after seeing the data.

2 Some Model Selection Approaches

To make things concrete, consider a binary classification setting, and let $\mathcal{H}_1, \mathcal{H}_2, \ldots \subseteq \{\pm 1\}^{\mathcal{X}}$ be a sequence of binary-valued function classes on $\mathcal{X}$. Given a training sample $S = ((x_1, y_1), \ldots, (x_m, y_m)) \in (\mathcal{X} \times \{\pm 1\})^m$, the goal is to select an index i_S and a function $h_S \in \mathcal{H}_{i_S}$ that performs well on future examples. Often, one assumes an underlying learning algorithm that for any given model index i, returns a function $h_S^i \in \mathcal{H}_i$ (e.g. consider using the SVM algorithm to learn a classifier in an RKHS corresponding to a polynomial kernel of degree i). In this case, the model selection algorithm can be viewed as a 'meta'-learning algorithm, whose goal is simply to select a model/function class index i_S (usually after taking into consideration the functions h_S^i returned by the learning algorithm for all function classes $\mathcal{H}_i$); the corresponding function $h_S^{i_S} \in \mathcal{H}_{i_S}$ is then returned.

Independent test (validation) sample. When labeled data is plenty and only a finite number of models are to be compared, a common model selection approach in practice is to withold some of the data to form an *independent test sample* $T = ((x_{m+1}, y_{m+1}), \ldots, (x_{m+n}, y_{m+n}))$, and to select the model i that gives the lowest error on this test sample; in other words to select $i_S = \arg\min_i \mathrm{er}_T[h_S^i]$. In this case, if $(S, T) \sim D_m \times D_n$ for some distribution D on $\mathcal{X} \times \{\pm 1\}$, then one can use Hoeffding's inequality together with the union bound to obtain a high confidence bound on $\mathrm{er}_D[h_S^{i_S}]$ in terms of $\mathrm{er}_T[h_S^{i_S}]$.

Cross-validation/leave-one-out. Another approach widely used in practice, in the absence of sufficient data to form a reliable holdout sample and as an approximation to the more ideal (but computationally intensive) *leave-one-out* method, is to use *k-fold cross-validation*, which involves dividing the m available examples S into k subsamples $S_1, \ldots, S_k$ (of roughly equal size) for some appropriate k (often $k = 5$ or 10), learning k functions $h^i_{S \setminus S_1}, \ldots, h^i_{S \setminus S_k} \in \mathcal{H}_i$ for each model i, where $S \setminus S_r$ denotes the sample formed

[1] Of course, one can argue this amounts to selecting a function from a function class consisting of the union of the individual function classes under consideration, which is true; however as we will see in this and the next lecture, allowing an algorithm to automatically select a class from a possibly infinite hierarchy of function classes can allow learning in a larger space overall than is possible with a fixed class of limited capacity, thereby leading to a potentially lower generalization error.

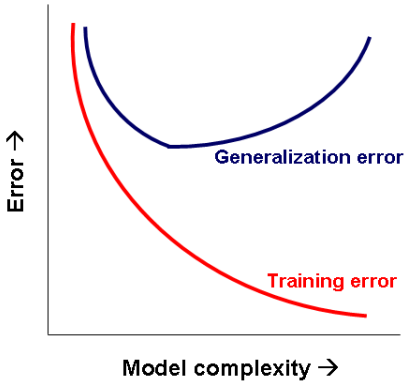

Figure 1: Typical picture of training and generalization errors as a function of increasing model complexity.

by removing S_r from S, testing each $h^i_{S\setminus S_r}$ on the left-out subsample S_r, and selecting the model i that gives lowest average error across the k folds (assuming again a finite number of models to be compared): $i_S = \arg\min_i \frac{1}{k}\sum_{r=1}^{k} \mathrm{er}_{S_r}[h^i_{S\setminus S_r}]$. The function $h^{i_S}_S \in \mathcal{H}_{i_S}$ is then returned. The leave-one-out method is an ideal case of this, with $k = m$, and in cases where generalization error bounds in terms of the leave-one-out error are available for each i, can be analyzed using methods similar to those we will discuss below. In certain cases, it is also possible to estimate such bounds without actually computing the leave-one-out error directly (e.g. see [2]).

Structural risk minimization. Our focus below will be on the situation when labeled data is limited and all available labeled examples are to be used in training. In this case, a popular model selection approach is that of *structural risk minimization* (SRM), which generalizes the empirical risk minimization (ERM) learning algorithm that given a function class $\mathcal{H}_i$, returns a function $h^i_S \in \mathcal{H}_i$ with minimal empirical risk (error) on the training sample S: $h^i_S = \arg\min_{h\in\mathcal{H}_i} \mathrm{er}_S[h]$ (for the zero-one loss, the minimum is always achieved; in general, one can allow for functions with error within some small precision of the infimum). SRM starts with a hierarchy of function classes $\mathcal{H}_1 \subset \mathcal{H}_2 \subset \ldots$ of increasing capacity or complexity, say $\mathrm{VCdim}(\mathcal{H}_1) < \mathrm{VCdim}(\mathcal{H}_2) < \ldots$, and uses ERM as the underlying learning algorithm to learn h^i_S within each H_i. Clearly, however, using the empirical risk $\mathrm{er}_S[h^i_S]$ alone as a criterion for model selection will simply select a model of high complexity, which may not lead to good generalization error (see Figure 1). SRM selects a function that minimizes the empirical risk plus a penalty term based on the complexity of the class (and the number of examples), for example $i_S = \arg\min_i \left(\mathrm{er}_S[h^i_S] + c\sqrt{\frac{\mathrm{VCdim}(\mathcal{H}_i)\ln m + i}{m}} \right)$, where $c > 0$ is some appropriate constant. More generally, similar ideas can be used when the functions $h^i_S \in \mathcal{H}_i$ are learned by an underlying learning algorithm other than ERM, namely, to add an appropriate model complexity term to the empirical error of each function, and to use this as a model selection criterion. These ideas are also related to the *minimum description length* (MDL) principle used in information-theoretic settings, and the more general approach of *complexity regularization*.

We will discuss properties of ERM and SRM in more detail in the next lecture; for now, we will focus on deriving generalization error bounds for the function learned by *any* model selection algorithm (including SRM), by obtaining bounds that hold uniformly for all functions $h \in \cup_i \mathcal{H}_i$. We will also see how these ideas can be extended to obtain, for example, margin-based bounds that hold uniformly over all values of the margin parameter $\gamma \in (0, 1]$ (thus allowing γ to be selected after observing the data).

Experimental and theoretical comparisons of various model selection algorithms, as well as pointers to several classical papers on the topic, can be found in [3, 1].

3 Generalization Error Bounds for Model Selection

As a simple example, let us start by considering two function classes $\mathcal{H}_1, \mathcal{H}_2$, say with $\mathrm{VCdim}(\mathcal{H}_1) < \mathrm{VCdim}(\mathcal{H}_2)$ (e.g. linear classifiers and quadratic classifiers). Suppose we use a learning algorithm to learn

a function $h_S^1 \in \mathcal{H}_1$ and a function $h_S^2 \in \mathcal{H}_2$, and observe that $\mathrm{er}_S[h_S^1] > \mathrm{er}_S[h_S^2]$. Which of h_S^1, h_S^2 should we prefer? Our goal of course is to minimize generalization error. We know that for each $i = 1, 2$, for any $0 < \delta \leq 1$, with probability at least $1 - \delta$ (over $S \sim D^m$):

$$\forall h \in \mathcal{H}_i : \quad \mathrm{er}_D[h] \;\leq\; \mathrm{er}_S[h] + c\sqrt{\frac{\mathrm{VCdim}(\mathcal{H}_i)\ln m + \ln(\frac{1}{\delta})}{m}}\,. \tag{1}$$

One approach therefore might be to pick $i_S \in \{1, 2\}$ for which the bound above is minimized; if the bound tracks accurately the true generalization error, this will lead us to the right model/function (this is the basic idea behind SRM).[2] Since we do not which of the two models will be selected, we can combine the above bound for $i = 1, 2$, taking each to hold with probability at least $1 - \delta/2$, to yield (by a simple application of the union bound) that with probability at least $1 - \delta$ (over $S \sim D^m$):

$$\forall i \in \{1, 2\}\ \forall h \in \mathcal{H}_i : \quad \mathrm{er}_D[h] \;\leq\; \mathrm{er}_S[h] + c\sqrt{\frac{\mathrm{VCdim}(\mathcal{H}_i)\ln m + \ln(\frac{2}{\delta})}{m}}\,. \tag{2}$$

Indeed, we could have chosen (before seeing the sample S) any $\delta_1, \delta_2 \in (0, 1]$ such that $\delta_1 + \delta_2 \leq \delta$ to get that with probability at least $1 - \delta$ (over $S \sim D^m$):

$$\forall i \in \{1, 2\}\ \forall h \in \mathcal{H}_i : \quad \mathrm{er}_D[h] \;\leq\; \mathrm{er}_S[h] + c\sqrt{\frac{\mathrm{VCdim}(\mathcal{H}_i)\ln m + \ln(\frac{1}{\delta_i})}{m}}\,. \tag{3}$$

Note that these bounds hold uniformly over *all* i and $h \in \mathcal{H}_i$, and therefore apply to any function learned from $\mathcal{H}_1 \cup \mathcal{H}_2$ regardless of the learning/model selection algorithm. The same approach is easily extended to any finite number of function classes $\mathcal{H}_1, \ldots, \mathcal{H}_k$; more generally, since the union bound can be applied to countable unions of events, we have the following:

Theorem 3.1. Let $\mathcal{H}_1, \mathcal{H}_2, \ldots \subseteq \{\pm 1\}^{\mathcal{X}}$ be a countable collection of function classes. Let $p_i \in [0, 1]$ such that $\sum_{i=1}^{\infty} p_i \leq 1$. Let D be any distribution on $\mathcal{X} \times \{\pm 1\}$. Then for any $0 < \delta \leq 1$, with probability at least $1 - \delta$ (over $S \sim D^m$):

$$\forall i \in \mathbb{N}\ \forall h \in \mathcal{H}_i : \quad \mathrm{er}_D[h] \;\leq\; \mathrm{er}_S[h] + c\sqrt{\frac{\mathrm{VCdim}(\mathcal{H}_i)\ln m + \ln(\frac{1}{p_i\delta})}{m}}\,.$$

Again note that this bound holds for *all* learning/model selection algorithms that pick a function from $\cup_i \mathcal{H}_i$. In practice, one might choose $p_i > 0$ for i such that $\mathrm{VCdim}(\mathcal{H}_i) < m$, and $p_i = 0$ for all other i, since in the latter case the bound is in any case not meaningful.

The above technique can also be used to obtain uniform bounds for real-valued function classes in terms of covering numbers or fat-shattering dimensions, or for both classification and regression using Rademacher averages. If specific bounds are available for the learning algorithm used within each function class (such as bounds based on algorithmic stability), then the above technique can also be used to obtain uniform versions of these bounds over the different function classes. We illustrate these ideas with a couple of examples.

Example 1. Consider using the SVM algorithm to learn a function

$$f_S^i = \arg\min_{f \in \mathcal{F}_K} \left(\mathrm{er}^{\mathrm{hinge}}[f] + \frac{\lambda_i}{2}\|f\|_K^2 \right) \tag{4}$$

for each i, where $\lambda_i > 0$ $(i = 1, 2, \ldots)$ are different regularization parameter values and $\mathcal{F}_K$ is the RKHS associated with a kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ with $K(x, x) \leq \kappa^2 < \infty\ \forall x \in \mathcal{X}$. We know from an earlier exercise that $f_S^i \in \mathcal{F}_i$, where

$$\mathcal{F}_i = \left\{ f \in \mathcal{F}_K \;\middle|\; \|f\|_K^2 \leq \frac{2}{\lambda_i} \right\}.$$

[2] Note though that bounds using fixed complexity terms such as the VC-dimension or fat-shattering dimension do not track accurately the generalization error for all distributions D; for this, distribution- or data-dependent complexity terms such as the Rademacher averages are typically more appropriate. See for example [3, 1, 4].

We also know that for any $x_1^m \in \mathcal{X}^m$, the empirical Rademacher average of $\mathcal{F}_i$ w.r.t. x_1^m is bounded as $R_{x_1^m}(\mathcal{F}_i) \leq \frac{1}{m}\sqrt{\frac{2}{\lambda_i}\sum_{j=1}^m K(x_j,x_j)}$. Moreover, by the reproducing kernel property, all functions $f \in \mathcal{F}_i$ satisfy $|f(x)| \leq \|f\|_K\sqrt{K(x,x)} \leq \kappa\sqrt{2/\lambda_i}$ for all $x \in \mathcal{X}$. If D is a probability distribution on $\mathcal{X} \times \{\pm 1\}$ and $\gamma > 0$ is a fixed parameter, then for each i, applying a previous result in terms the empirical Rademacher averages of $\mathcal{F}_i$ then gives that for any $0 < \delta \leq 1$, with probability at least $1-\delta$ (over $S \sim D^m$),

$$\forall f \in \mathcal{F}_i : \quad \mathrm{er}_D[f] \;\leq\; \mathrm{er}_S^{\mathrm{ramp}(\gamma)}[f] + \frac{2}{\gamma m}\sqrt{\frac{2}{\lambda_i}\sum_{j=1}^m K(x_j,x_j)} + 2\sqrt{\frac{\ln(\frac{4}{\delta})}{2m}} + 2\kappa\sqrt{\frac{\ln(\frac{2}{\delta})}{\lambda_i m}}\,. \tag{5}$$

Applying the above technique then gives that for any $p_i \in [0,1]$ with $\sum_{i=1}^\infty p_i \leq 1$ and any $0 < \delta \leq 1$, with probability at least $1-\delta$ (over $S \sim D^m$),

$$\forall i \in \mathbb{N}\; \forall f \in \mathcal{F}_i : \quad \mathrm{er}_D[f] \;\leq\; \mathrm{er}_S^{\mathrm{ramp}(\gamma)}[f] + \frac{2}{\gamma m}\sqrt{\frac{2}{\lambda_i}\sum_{j=1}^m K(x_j,x_j)} + 2\sqrt{\frac{\ln(\frac{4}{p_i\delta})}{2m}} + 2\kappa\sqrt{\frac{\ln(\frac{2}{p_i\delta})}{\lambda_i m}}\,. \tag{6}$$

Note that this bound is uniform over both $i \in \mathbb{N}$ and $f \in \mathcal{F}_i$, and therefore applies to *any* function learned from $\cup_i \mathcal{F}_i$, not just those learned by the SVM algorithm.

Example 2. Let $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^n \mid \|\mathbf{x}\|_2 \leq 1\}$, and consider now using the SVM algorithm to learn a function

$$f_S^i = \arg\min_{f \in \mathcal{F}_i}\left(\mathrm{er}^{\mathrm{hinge}}[f] + \frac{\lambda}{2}\|f\|_{K_i}^2\right) \tag{7}$$

for each i, where $K_i(\mathbf{x},\mathbf{x}') = (1+\mathbf{x}\cdot\mathbf{x}')^i$, $\mathcal{F}_i = \mathcal{F}_{K_i}$, and $\lambda > 0$ is a fixed regularization parameter. Note that $K_i(\mathbf{x},\mathbf{x}) \leq 2^i\ \forall \mathbf{x} \in \mathcal{X}$. If D is a probability distribution on $\mathcal{X} \times \{\pm 1\}$ and $\gamma > 0$ is a fixed parameter, then for each i, applying the algorithmic stability results of the last lecture, we have that for any $0 < \delta \leq 1$, with probability at least $1-\delta$ (over $S \sim D^m$),

$$\mathrm{er}_D[f_S^i] \;\leq\; \mathrm{er}_S^{\mathrm{ramp}(\gamma)}[f_S^i] + \frac{4\cdot 2^i}{\gamma\lambda m} + \left(\frac{8\cdot 2^i}{\gamma\lambda} + 1\right)\sqrt{\frac{\ln(\frac{1}{\delta})}{2m}}\,. \tag{8}$$

Applying the above technique then gives that for any $p_i \in [0,1]$ with $\sum_{i=1}^\infty p_i \leq 1$ and any $0 < \delta \leq 1$, with probability at least $1-\delta$ (over $S \sim D^m$),

$$\forall i \in \mathbb{N} : \quad \mathrm{er}_D[f_S^i] \;\leq\; \mathrm{er}_S^{\mathrm{ramp}(\gamma)}[f_S^i] + \frac{4\cdot 2^i}{\gamma\lambda m} + \left(\frac{8\cdot 2^i}{\gamma\lambda} + 1\right)\sqrt{\frac{\ln(\frac{1}{p_i\delta})}{2m}}\,. \tag{9}$$

Note that in this case the bound is not uniform over all $f \in \mathcal{F}_i$; it is uniform over $i \in \mathbb{N}$, but for each i, holds for the specific function f_S^i learned by the SVM algorithm using K_i.

4 Uniform Margin Bounds

Let $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}}$. Let D be any probability distribution on $\mathcal{X} \times \{\pm 1\}$. Recall the following margin-based generalization error bound we have derived in an earlier lecture, which applies to binary classification algorithms that learn a real-valued function $f_S \in \mathcal{F}$ and then classify according to $\mathrm{sign}(f_S)$: for any fixed $\gamma > 0$ and any $0 < \delta \leq 1$, with probability at least $1-\delta$ (over $S \sim D^m$),

$$\forall f \in \mathcal{F} : \quad \mathrm{er}_D^{0\text{-}1}[f] \;\leq\; \mathrm{er}_S^{\mathrm{margin}(\gamma)}[f] + c\sqrt{\frac{\ln \mathcal{N}_\infty(\frac{\gamma}{2},\mathcal{F},2m) + \ln(\frac{1}{\delta})}{m}}\,, \tag{10}$$

where $c > 0$ is some fixed constant. While useful, this bound is limited by the fact that the margin parameter γ must be chosen *before* seeing the data; in practice, we would like to be able to apply such a bound using a value for γ that depends on the sample S. One way to achieve this is via a uniform bound over γ. This is slightly more tricky since γ is a continuous parameter, so we cannot apply the union bound in a straightforward manner as done above for model selection over a countably infinite number of function classes $\mathcal{H}_i$. However an adaptation of the above technique that divides any bounded interval for γ into a countably infinite union of intervals allows us to obtain such a uniform margin bound:

Theorem 4.1 (Bartlett, 1998)**.** Let $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}}$. Let D be any probability distribution on $\mathcal{X} \times \{\pm 1\}$. Then for any $0 < \delta \leq 1$, with probability at least $1 - \delta$ (over $S \sim D^m$),

$$\forall \gamma \in (0,1] \;\; \forall f \in \mathcal{F} : \;\; \mathrm{er}_D^{0\text{-}1}[f] \;\leq\; \mathrm{er}_S^{\mathrm{margin}(\gamma)}[f] + c\sqrt{\frac{\ln \mathcal{N}_\infty(\frac{\gamma}{4}, \mathcal{F}, 2m) + \ln(\frac{2}{\delta\gamma})}{m}}\,.$$

Proof. For any $\gamma_1, \gamma_2, \delta \in (0,1]$, define the event

$$E(\gamma_1, \gamma_2, \delta) = \left\{ \exists f \in \mathcal{F} : \mathrm{er}_D^{0\text{-}1}[f] > \mathrm{er}_S^{\mathrm{margin}(\gamma_1)}[f] + c\sqrt{\frac{\ln \mathcal{N}_\infty(\frac{\gamma_2}{2}, \mathcal{F}, 2m) + \ln(\frac{1}{\delta})}{m}} \right\}.$$

Now,

$$\begin{aligned}
&\mathbf{P}_{S\sim D^m}\left(\exists \gamma \in (0,1], f \in \mathcal{F} : \mathrm{er}_D^{0\text{-}1}[f] > \mathrm{er}_S^{\mathrm{margin}(\gamma)}[f] + c\sqrt{\frac{\ln \mathcal{N}_\infty(\frac{\gamma}{4}, \mathcal{F}, 2m) + \ln(\frac{2}{\delta\gamma})}{m}}\right) && (11)\\
&= \mathbf{P}_{S\sim D^m}\left(\exists \gamma \in (0,1] : E\left(\gamma, \frac{\gamma}{2}, \frac{\delta\gamma}{2}\right)\right) && (12)\\
&= \mathbf{P}_{S\sim D^m}\left(\bigcup_{i=1}^{\infty}\left\{\exists \gamma \in \left(\frac{1}{2^{i+1}}, \frac{1}{2^i}\right] : E\left(\gamma, \frac{\gamma}{2}, \frac{\delta\gamma}{2}\right)\right\}\right) && (13)\\
&\leq \sum_{i=1}^{\infty} \mathbf{P}_{S\sim D^m}\left(\exists \gamma \in \left(\frac{1}{2^{i+1}}, \frac{1}{2^i}\right] : E\left(\gamma, \frac{\gamma}{2}, \frac{\delta\gamma}{2}\right)\right) \quad \text{(by union bound)} && (14)\\
&\leq \sum_{i=1}^{\infty} \mathbf{P}_{S\sim D^m}\left(E\left(\frac{1}{2^{i+1}}, \frac{1}{2^{i+1}}, \frac{\delta}{2^{i+1}}\right)\right) && (15)\\
&\qquad\qquad \text{(since } E(\gamma_1,\gamma_2,\delta) \implies E(\gamma_1',\gamma_2',\delta') \;\; \forall \gamma_1' \leq \gamma_1, \gamma_2' \geq \gamma_2, \delta' \geq \delta). && (16)\\
&\leq \sum_{i=1}^{\infty} \frac{\delta}{2^{i+1}} \quad \text{(by Eq. (10), applied with } \gamma = \tfrac{1}{2^{i+1}} \text{ and confidence parameter } \tfrac{\delta}{2^{i+1}}) && (17)\\
&= \delta\,. && (18)
\end{aligned}$$

□

Exercise. Apply the above technique to obtain a uniform bound over $\gamma \in (0,1]$ for the 0-1 generalization error $\mathrm{er}_D^{0\text{-}1}[f_S]$ of a function f_S learned from a function class $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}}$ in terms of the empirical $\ell_{\mathrm{ramp}(\gamma)}$ error $\mathrm{er}_S^{\mathrm{ramp}(\gamma)}[f_S]$ using the Rademacher averages of $\mathcal{F}$ (use results from Lecture 7).

Exercise. Apply the above technique to obtain a uniform bound over $\gamma \in (0,1]$ for the 0-1 generalization error $\mathrm{er}_D^{0\text{-}1}[f_S]$ of a function $f_S : \mathcal{X} \to \mathbb{R}$ learned by an algorithm with uniform score stability $\nu : \mathbb{N} \to [0,\infty)$ in terms of the empirical $\ell_{\mathrm{ramp}(\gamma)}$ error $\mathrm{er}_S^{\mathrm{ramp}(\gamma)}[f_S]$ (use results from the previous lecture).

5 Next Lecture

In the next lecture, we will start to shift gears from the finite sample setting we have focused on so far to the infinite sample limit, where we consider the behaviour of learning algorithms in the limit of infinite sample size. We will define the excess error, approximation error, and estimation error which will be useful in this context, and will start discussing the notions of statistical consistency and learnability. We will also study statistical consistency properties of ERM and SRM, and will find that the tools we have studied so far will be useful for this purpose as well.

References

[1] Peter L. Bartlett, Stephane Boucheron, and Gabor Lugosi. Model selection and error estimation. *Machine Learning*, 48:85–113, 2002.

[2] Thorsten Joachims. Estimating the generalization performance of an SVM efficiently. In *Proceedings of the 17th International Conference on Machine Learning (ICML)*, 2000.

[3] Michael Kearns, Yishay Mansour, Andrew Y. Ng, and Dana Ron. An experimental and theoretical comparison of model selection methods. *Machine Learning*, 27:7–50, 1997.

[4] Vladimir Koltchinskii. Rademacher penalties and structural risk minimization. *IEEE Transactions on Information Theory*, 47(5):1902–1914, 2001.