Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods

Paper Authors: John C Platt

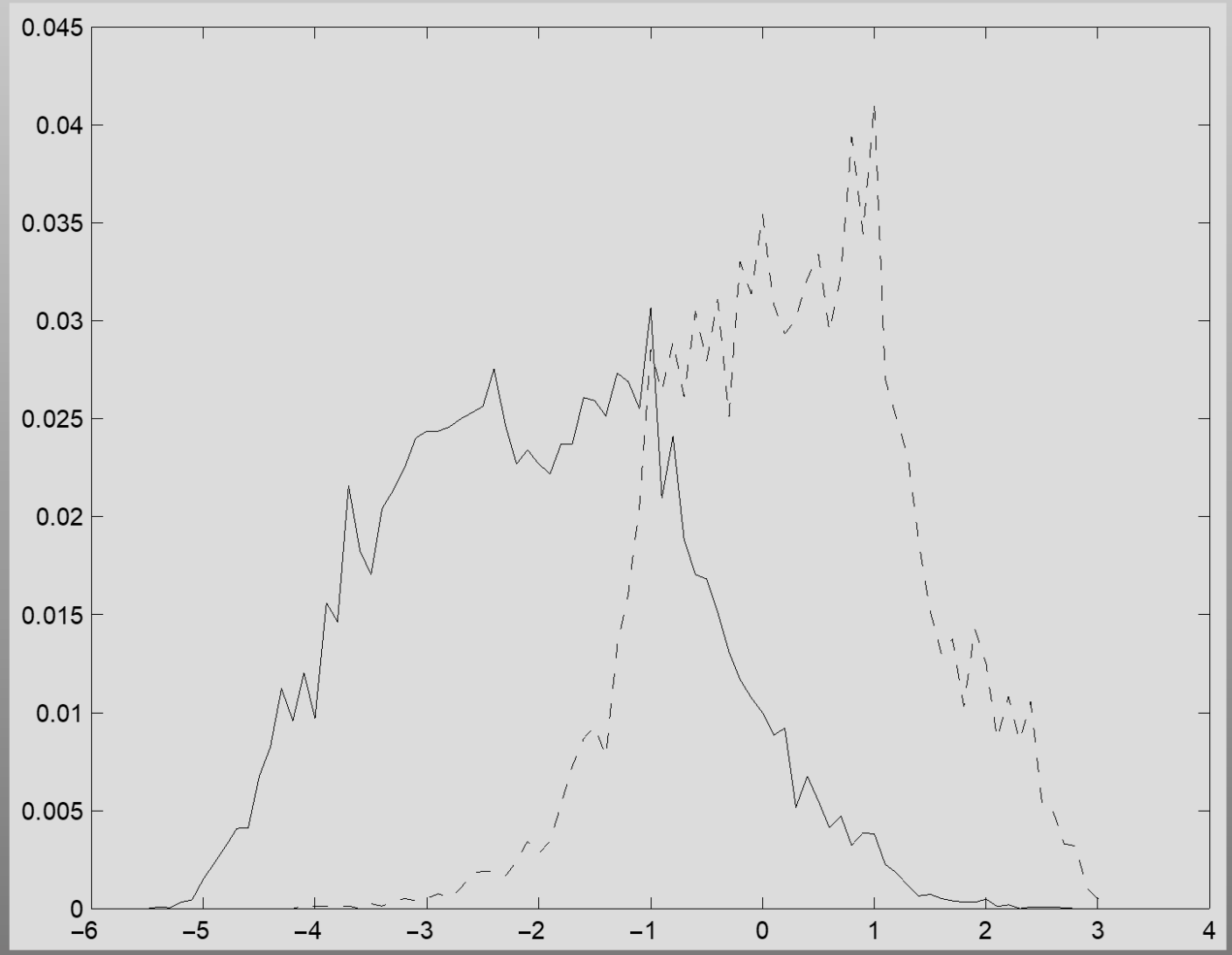

Image reproduced under fair use from https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.41.1639

Overview

- Desirable output of a classifier is a calibrated posterior probability
  - Facilitates post-processing.
- Standard SVMs do not provide such probabilities.
- Create probabilities by training a kernel classifier with
    - logit link function and
    - a regularized maximum likelihood score.
- Training with a maximum likelihood score produces non-sparse kernels.
- Instead,
  - train an SVM,
  - train the parameters of a sigmoid function that
    - maps the SVM outputs into probabilities.
- The SVM + sigmoid comparable to the regularized maximum likelihood kernel
  - Retains the sparseness of the SVM.

Introduction

- Construct a classifier to produce a posterior probability P(class I input)
  - Allows decisions that use a utility model
  - Important when a classifier is making a small part of an overall decision
  - Combine different classifier outputs
    - Viterbi search or HMM: results from phoneme recognizers into word recognition.
  - Multi-label classifier:
    - label with maximal posterior probability is Bayes optimal for equal loss case
- SVMs produce an uncalibrated value that is not a probability
- The unthresholded output of an SVM: $f(x) = h(x) + b$, where

$$h(\mathbf{x}) = \sum y_i \alpha_i k(\mathbf{x}_i, \mathbf{x})$$

- Training minimizes

$$C \sum (1 - y_i f_i)_+ + \frac{1}{2} ||h||_{\mathcal{F}},$$

Related Work - I

- Logistic link function by Wahba

$$P(\text{class}|\text{input}) = P(y = 1|\mathbf{x}) = p(\mathbf{x}) = \frac{1}{1 + \exp(-f(\mathbf{x}))},$$

- Minimize a negative log multinomial likelihood
  - plus a term that penalizes the norm

$$-\frac{1}{m}\sum_i \left(\frac{y_i + 1}{2} \log(p_i) + \frac{1 - y_i}{2} \log(1 - p_i) \right) + \lambda ||h||^2_{\mathcal{F}},$$

Wahba G. Support vector machines, reproducing kernel Hilbert spaces and the randomized GACV. Advances in Kernel Methods-Support Vector Learning. 1999 Feb 8;6:69-87.

$$C \sum (1 - y_i f_i)_+ + \frac{1}{2}||h||_{\mathcal{F}},$$

Related Work - II

- Logistic Link by Wahba

$$P(\text{class}|\text{input}) = P(y = 1|\mathbf{x}) = p(\mathbf{x}) = \frac{1}{1 + \exp(-f(\mathbf{x}))},$$

- Output $p(x)$ of such a machine is a posterior probability.
- Minimizing this error function will not directly produce a sparse SVM

$$-\frac{1}{m}\sum_i \left(\frac{y_i + 1}{2}\log(p_i) + \frac{1 - y_i}{2}\log(1 - p_i)\right) + \lambda||h||^2_{\mathcal{F}}$$

- But a modification can produce sparse kernel machines

Wahba G, Lin X, Gao F, Xiang D, Klein R, Klein B. The bias-variance tradeoff and the randomized GACV. Advances in Neural Information Processing Systems. 1998;11.

Related Work - III

- Map SVM output to probabilities by decomposing feature space $\mathcal{F}$
  - a direction orthogonal to the separating hyperplane,
  - and all of the $N - 1$ other dimensions of the feature space.
- The orthogonal direction is parameterized by t
  - A scaled version of $f(x)$
- All other directions parameterized by a vector **u.**
- In general, the posterior depends on both t and $\boldsymbol{u}$: $P(y = 1|\, t,\ \boldsymbol{u})$.
  - Vapnik fits this probability with a sum of cosine terms with strong results

Vapnik V. The nature of statistical learning theory. Springer science & business media; 1999 Nov 19.

$$P(y = 1|t, \mathbf{u}) = a_0(\mathbf{u}) + \sum_{n=1}^{N} a_n(\mathbf{u}) \cos(nt).$$

- Requires a solution of a linear system for every evaluation of the SVM.
  - The Platt scaling approach avoids it.
  - The dependencies of $P(y = 1|f)$ on $\boldsymbol{u}$ cannot be modeled.

Related Work - IV

- Fit Gaussians to the class-conditional densities of the SVM outputs
  - $p(f|y = \mathbf{1})$ and
  - $p(f|y = -1)$
- A single tied variance is estimated for both Gaussians.
- The posterior probability rule $P(y = 1|f)$ is thus a sigmoid
  - slope determined by the variance.
- Adjust the bias of the sigmoid
  - such that the point $P(y = 1|f) = 0.5$ occurs at $f = 0$.
- The single parameter may not model the true posterior probability.

Hastie T, Tibshirani R. Classification by pairwise coupling. Advances in neural information processing systems. 1997;10.

Related Work - V

- Employ a more flexible version of the Gaussian fit to $p(f \mid y = \pm 1)$
- Mean and variance for each Gaussian is determined from a data set
- Bayes' rule

$$P(y = 1 \mid f) = \frac{p(f \mid y = 1)P(y = 1)}{\sum_{i=-1,1} p(f \mid y = i)P(y = i)},$$

  - $P(y = i)$: prior probabilities computed from the training set
- This model for SVM output probabilities independently proposed
  - Used for speaker identification by C. J. C. Burges at 1998 NIPS SVM workshop
- The posterior is an analytic function of f with form:

$$P(y = 1 \mid f) = \frac{1}{1 + \exp(af^2 + bf + c)},$$

Related Work - VI

- Two issues with this approach:
  - the assumption of Gaussian class-conditional densities is often violated
  - the posterior estimate derived from the two-Gaussian approximation is non-monotonic

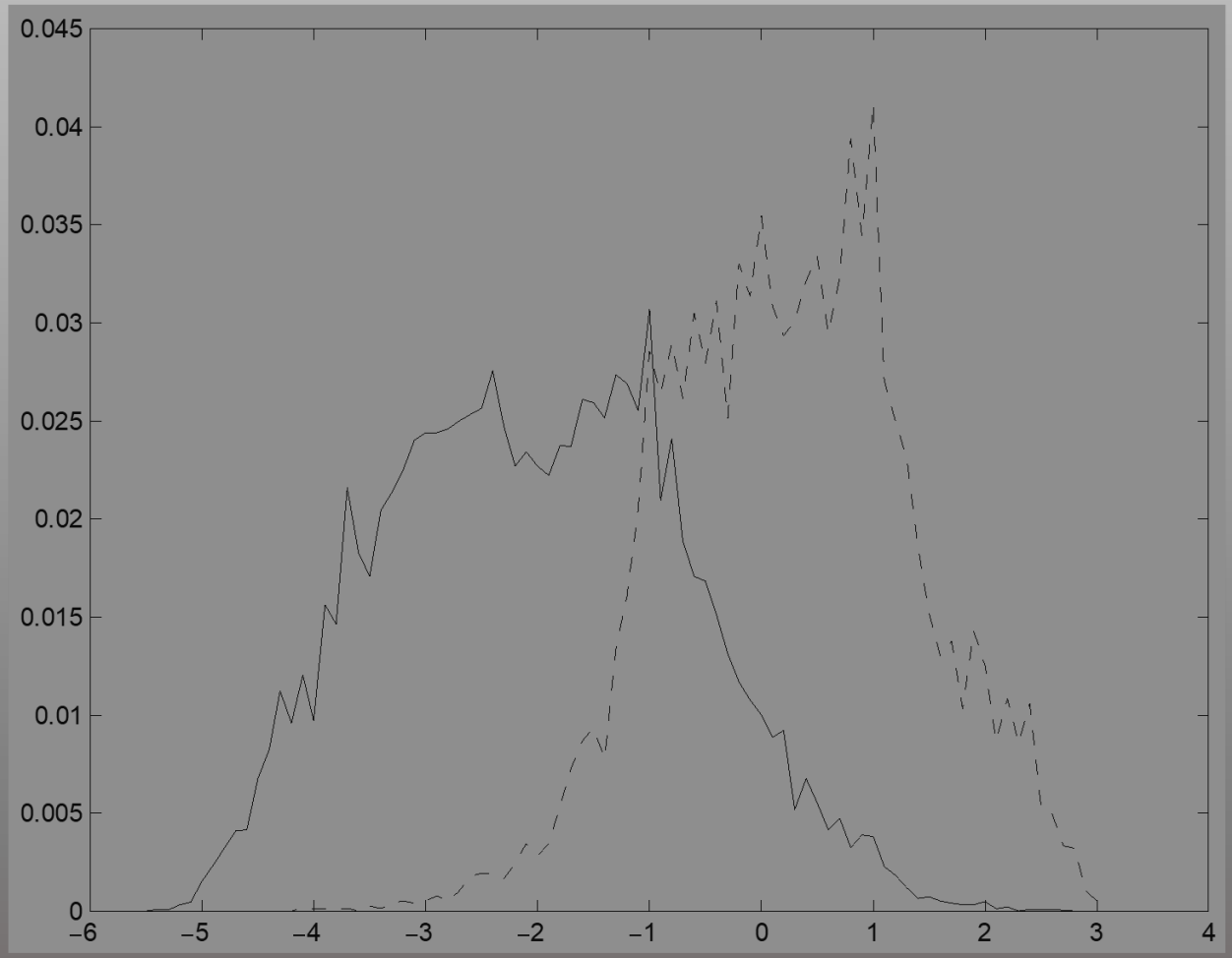

Image reproduced under fair use from https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.41.1639

The Platt Scaling Approach

- Use a parametric model to fit the posterior $P(y = 1 \mid f)$ directly
  - instead of estimating the class-conditional densities $p(f \mid y)$
- The parameters can be adapted to give the best probability outputs
- The form of the parametric model inspired by empirical data
  - Far away from Gaussian

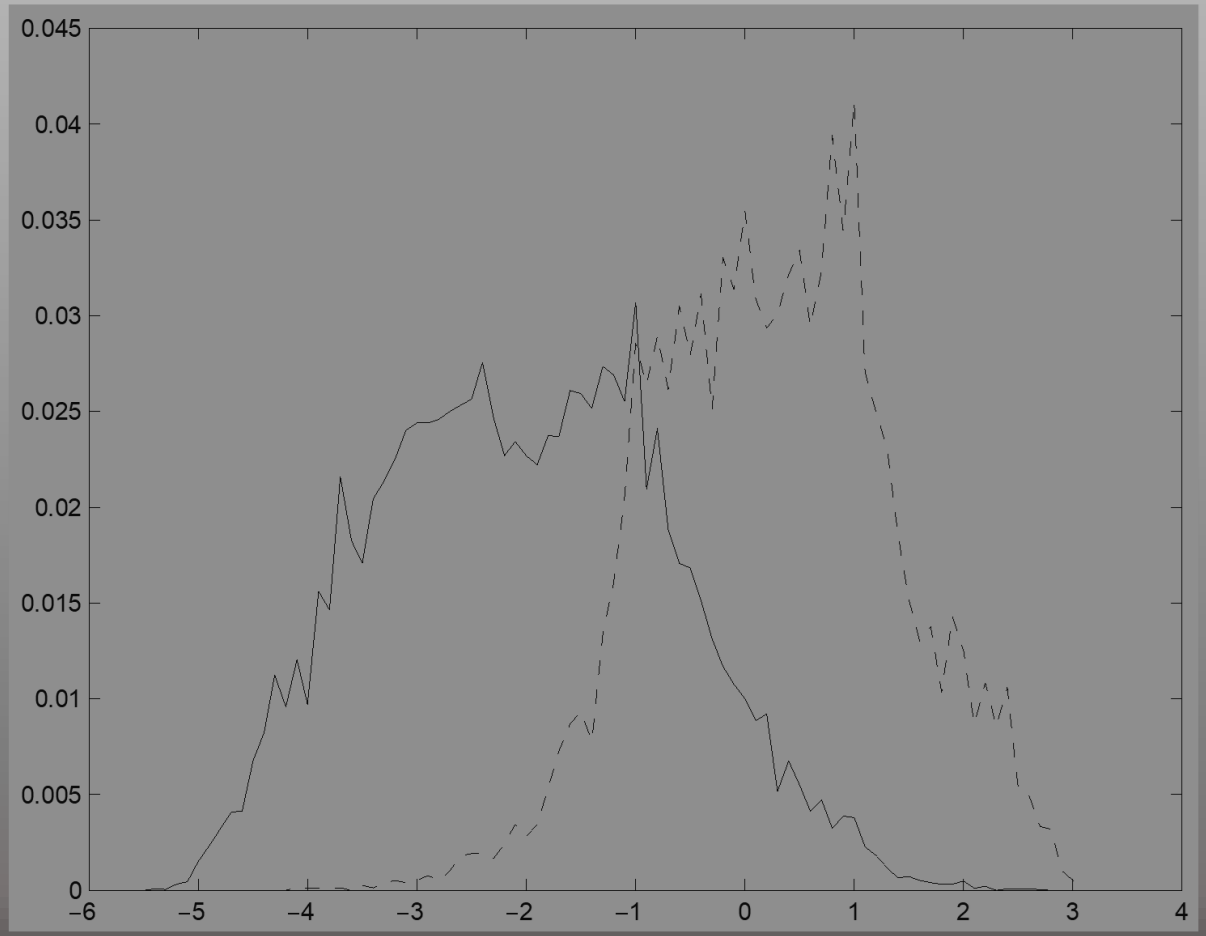

Image reproduced under fair use from https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.41.1639

Approach - II

- The class-conditional densities apparently exponential.
- Bayes' rule on two exponentials suggests using a parametric form of a sigmoid:

$$P(y = 1 \mid f) = \frac{1}{1 + \exp(Af + B)}$$

- Equivalent to assuming that SVM output proportional to log odds of a positive example.

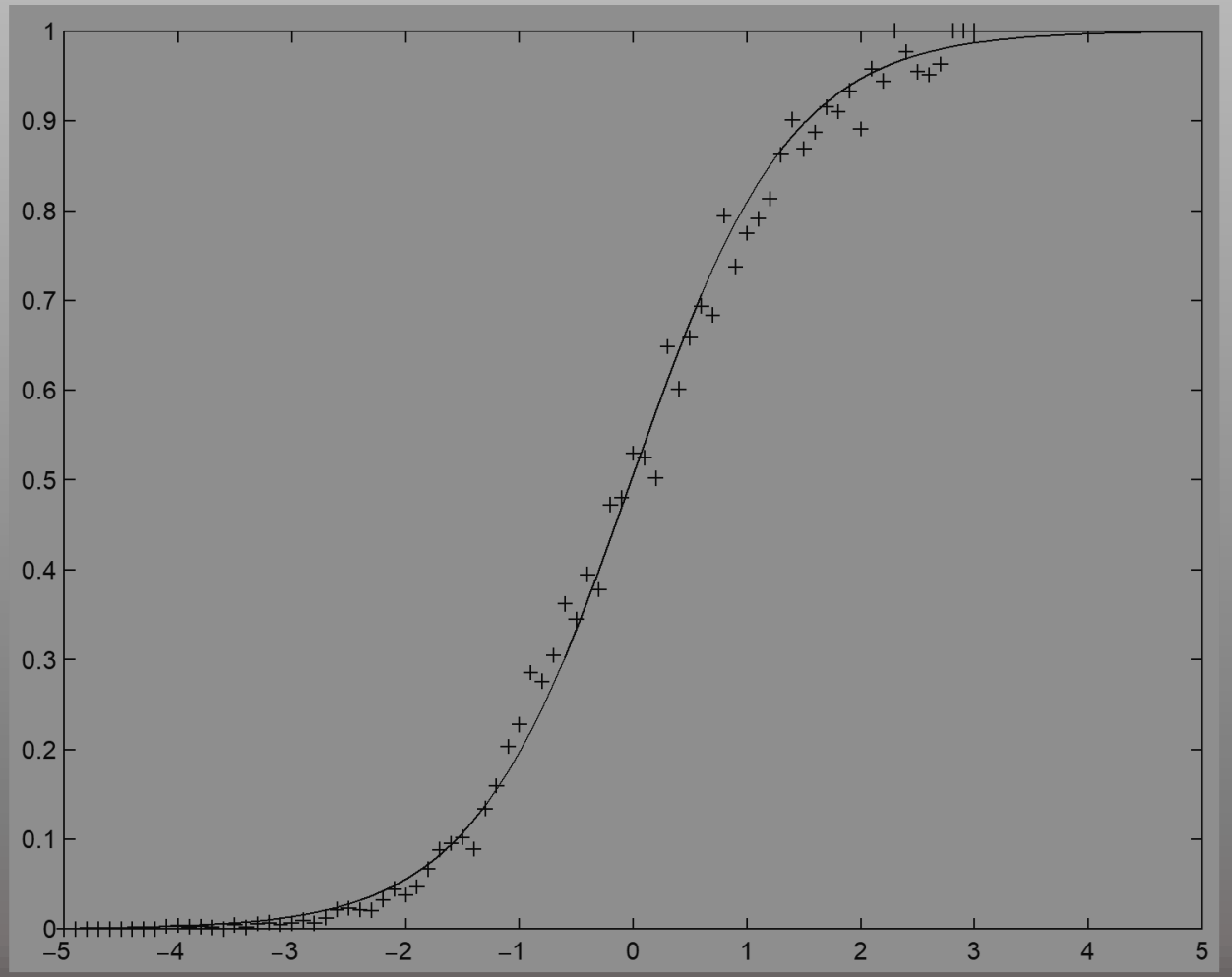

Image reproduced under fair use from https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.41.1639

Approach - III

$$P(y = 1 \mid f) = \frac{1}{1 + \exp(Af + B)}.$$

- Parameters A and B fitted using maximum likelihood estimation
- Training set (f_i, y_i)
- Step 1: Define a new training set (f_i, t_i) where t_i are target probabilities:

$$t_i = \frac{y_i + 1}{2}.$$

- Minimize the negative log likelihood or cross-entropy of the training data:

$$\min -\sum_i t_i \log(p_i) + (1 - t_i) \log(1 - p_i),$$

Approach - IV

- Two practical challenges
  - How to obtain training data t_i?
  - How to avoid overfitting?
- Using all of the training data is not a good idea
  - At the margins, $f(x_i) = 1$
  - Not likely to be true of any test data
  - So, test data will certainly become OOD.
- Solutions:
  - Estimate $f(x_i)$ by performing leave-one-out estimation
    - Expensive

Approach - V

- Using all of the training data is not a good idea
  - test data will certainly become OOD.
- Solutions:
  - Estimate $f(x_i)$ by performing leave-one-out estimation
  - Hold-out set
    - Do not use 30% of data to train the SVM
    - Use this data to train the sigmoid
    - Needs more data
  - Cross-validation
    - Split data into three parts
    - Train on permutation of 2 and fit the sigmoid on the third
    - Can be extended to n-fold cross validation

Results - I

Experiment 1

- Assuming equal loss for Type I/II errors,
- Optimal threshold for SVM+sigmoid is
  - $P(y = 1 \mid f) = 0.5$
- Optimal threshold for SVM is
  - $f = 0$
- Achieved. f = -0.17 in experiments.

Experiment 2

- Compare SVM+sigmoid to an explicit approach
  - that maximizes log multinomial likelihood

Results - II

Task	Raw SVM Number of Errors	SVM + Sigmoid Number of Errors	Regularized Likelihood Number of Errors	SVM + Sigmoid $-\log(p)$ Score	Regularized Likelihood $-\log(p)$ Score
Reuters Linear	1043	963	1060	3249	3301
Adult Linear	2441	2442	2434	5323	5288
Adult Quadratic	2626	2554	2610	5772	5827
Web Linear	260	265	248	1121	958
Web Quadratic	444	452	507	1767	2163

- Adding a Sigmoid often helps the SVM!
- Neither approach (Sigmoid or regularized likelihood) is better.
- Adding sigmoid produces probabilities comparable to regularized likelihood.

Conclusions

- Method for extracting probabilities P(class | input) from SVM outputs
- Does not alter the training of the SVM
  - No regularization term
- Trainable post-processing step with a binomial maximum likelihood
- Two-parameter sigmoid used for post-processing
  - As it is observed empirically
- SVM + sigmoid comparable in accuracy to SVMs
  - Or better
- SVM + sigmoid preserves sparseness of kernels
- SVM + sigmoid produces probabilities comparable to regularization.