Probabilistic Score Estimation with Piecewise Logistic Regression

Jian Zhang JIAN.ZHANG@CS.CMU.EDU
Yiming Yang YIMING@CS.CMU.EDU
School of Computer Science, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213 USA

Abstract

Well-calibrated probabilities are necessary in many applications like probabilistic frameworks or cost-sensitive tasks. Based on previous success of asymmetric Laplace method in calibrating text classifiers' scores, we propose to use piecewise logistic regression, which is a simple extension of standard logistic regression, as an alternative method in the discriminative family. We show that both methods have the flexibility to be piecewise linear functions in log-odds, but they are based on quite different assumptions. We evaluated asymmetric Laplace method, piecewise logistic regression and standard logistic regression over standard text categorization collections (Reuters-21578 and TREC-AP) with three classifiers (SVM, Naive Bayes and Logistic Regression Classifier), and observed that piecewise logistic regression performs significantly better than the other two methods in the log-loss metric.

1. Introduction

Among existing popular classifiers only a few of them are able to generate posterior probabilities. It is very useful, on the other hand, to have system estimated probabilities along with classification decisions. For example, probabilities can be directly plugged into hierarchical classification schemes where non-deterministic decisions are required, or in cost-sensitive tasks or utility models (Duda et. al, 2001) where expected risk/utility is needed. Probability estimates can also be used for uncertainty sampling in active learning (Lewis & Gale, 1994).

It has been well understood that classification tasks can be approached from either generative models where the class-conditional densities are learned, or discriminative models where decision boundaries are directly learned regardless of the class-conditional densities. It is generally observed that discriminative methods perform better than generative ones in terms of classification tasks, and the strength and weakness of both groups have been discussed and analyzed in the literature (Rubinstein & Hastie, 1997; Ng & Jordan, 2002). For the task of probabilistic score estimation (score calibration), logistic regression (Platt, 1999) has been one of the most widely used methods in the discriminative family. Recently, it has been shown (Bennett, 2003) that asymmetric Laplace method, which is a generative model that allows more flexible class-conditional densities than Laplace distribution, outperforms logistic regression in calibrating text classifiers' output scores. It would be interesting to investigate whether logistic regression can be further improved to achieve comparable results in this task.

In this paper we propose to use piecewise logistic regression, a simple extension of standard logistic regression, as an alternative approach to the asymmetric Laplace method for score calibration. Belonging to the discriminative family, piecewise logistic regression needs fewer assumptions than generative models like the asymmetric Laplace method, but also has the same flexibility as being piecewise linear in log-odds as the asymmetric Laplace method. And regularization is added to avoid overfitting in local nosiy region. We analyze and compare these two methods for a better understanding, and examine them as well as standard logistic regression on several text categorization benchmark collections with three classifiers.

The rest of the paper is organized as follows. In Section 2, we review related work on estimating probabilistic scores from both generative and discriminative viewpoints. In Section 3, we first introduce two methods used in probability estimation, namely logistic regression and asymmetric Laplace, then propose our piecewise logistic regression method. We also compare and

Appearing in *Proceedings of the 21st International Conference on Machine Learning*, Banff, Canada, 2004.

analyze asymmetric Laplace and piecewise logistic regression, and discusse more general approaches to the problem. In Section 4, we describe the experimental design, and in Section 5 we present experimental results comparing three calibration methods. Finally, we conclude our work in Section 6.

2. Related Work

For generative models, class priors and class-conditional probabilities should be first obtained, and then Bayes rule is applied to get the posterior probabilities; on the other hand, discriminative models directly compute posterior probabilities without considering class-conditional probabilities.

Among discriminative models, logistic regression has been one of the most frequently used methods for estimating probabilities. Platt (1999) used logistic regression to convert the output scores of SVM to probabilistic scores and showed that the calibrated scores yield comparable probabilistic results to regularized logistic regression classifier. Zadrozny and Elkan (2002) proposed isotonic regression which is non-decreasing stepwise-constant to model posterior probabilities, and applied their method to Naive Bayes and SVM. However, their resulting estimation curve is discontinuous and the cost to store the model is high.

Among generative models for estimating probabilities, Gaussian distributions have been used (Hastie & Tibshirani, 1996) to fit class-conditional densities. There are two variants of this approach: assume both Gaussians have the same variance and fit means and the common variance; or fit both means and variances. As discussed in previous work (Platt, 1999), the former candidate has the disadvantage of oversimplifying the model, while the latter essentially violates the monotonicity property due to its quadratic term in the exponential function. Based on the observation from information retrieval that score distributions for "extremely irrelevant", "hard to discriminate" and "obviously relevant" documents are often quite different, Bennett (2003) proposed to use asymmetric distributions to model the scores of both classes, and showed that asymmetric Laplace distribution gives better performance than logistic regression in calibrating text classifiers' outputs. However, his comparison between asymmetric distributions and logistic regression is "unfair" in the sense that the former allows piecewise linear fit in log-odds while the latter is restricted to a global linear fit. A better examination, in our opinion, would be comparing both methods in the setting of piecewise linear fit, and this is what we are going to investigate in this paper.

3. Methods

The probability estimation problem is defined as follows. A training set $D = \{(x_1, y_2), (x_2, y_2), \ldots, (x_N, y_N)\}$ is given to the system, where $x_1, x_2, \ldots, x_N$ is a list of scores generated by some classifier $\mathbf{C}$ over instances $\mathbf{X} = \{\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_N\}$ and $y_1, y_2, \ldots, y_N$ are the corresponding class labels for instances in $\mathbf{X}$. Throughout the paper we assume that there are two classes with $y = 1$ or '+' for the positive class and $y = -1$ or '-' for the negative class, unless otherwise specified.

3.1. Logistic Regression

Logistic regression is one of the most frequently used discriminative methods to the probability estimation problem (Platt, 1999; Bennett, 2003; Lewis & Gale, 1994). The standard logistic regression assumes that the posterior probability has the form of a sigmoid function $P(y|x) = \frac{1}{1+\exp(-y(ax+b))}$, which is equivalent to say that the log-odds is a linear function in x: $\log \frac{P(+|x)}{P(-|x)} = ax + b$.

Regularization can be added to avoid overfitting. The parameters of regularized logistic regression can be estimated with maximum likelihood estimation using a training set $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$, which is equivalent to minimize the following quantity:

$$\mathbf{O} = \sum_{i=1}^{N} \log(1 + \exp(-y_i(ax_i + b))) + \lambda(a^2 + b^2)$$

where λ is the regularization coefficient that controls the balance between training loss and model complexity.

3.2. Asymmetric Laplace Method

Unlike standard Laplace distribution, asymmetric Laplace distribution allows different slopes around the mode. It has been shown that asymmetric Laplace method achieved good performance in calibrating text classifiers' outputs (Bennett, 2003). The class-conditional density of asymmetric Laplace method is given by:

$$p(x|\theta, \beta, \gamma) = \begin{cases} \frac{\beta\gamma}{\beta+\gamma} \exp[-\beta(\theta - x)] & x \leq \theta \\ \frac{\beta\gamma}{\beta+\gamma} \exp[-\gamma(x - \theta)] & x > \theta \end{cases}$$

where $\beta, \gamma > 0$ and θ are model parameters.

It degenerates to standard Laplace distribution when $\beta = \gamma$. In Figure 1 we show the asymmetric Laplace distributions with $\beta = 1.0$ and $\gamma = 0.5, 1.0, 2.0, 5.0$. From the graph we can see that by allowing different slopes on both sides of the mode, asymmetric Laplace distributions are able to fit a more general class of distributions.

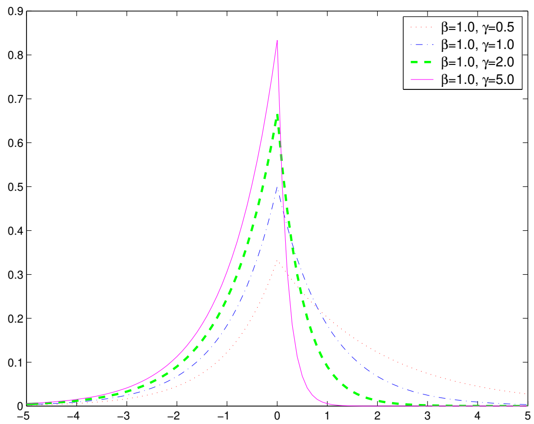

Figure 1. Asymmetric Laplace Distributions

For a fixed θ, the Maximum Likelihood Estimation (MLE) of model parameters β and γ can be computed directly from the training data as (Bennett, 2003): $\beta_{MLE} = \frac{N}{X_l+\sqrt{X_l X_r}}$ and $\gamma_{MLE} = \frac{N}{X_r+\sqrt{X_r X_l}}$ where $X_l = \theta|\{x \in X | x \leq \theta\}| - \sum_{x\in D, x\leq\theta} x$ and $X_r = \sum_{x\in D, x>\theta} x - \theta \cdot |\{x \in D | x > \theta\}|$.

And class priors are estimated with smoothed add-one estimator $P(y) = \frac{N_y+1}{N+2}$ where N is the total number of examples in the training set and N_y is the number of training examples that belong to class y. We follow the method used in (Bennett, 2003) to choose θ: loop over individual x_i and set θ to the one which gives maximum likelihood[1].

3.3. Piecewise Logistic Regression

By its name, piecewise logistic regression is a simple extension of the standard logistic regression by replacing the linear function into a piecewise linear function in the exponential term. Correspondingly, the log-odds of the piecewise logistic regression is also piecewise linear, which gives much more flexibility than a global linear function.

The intuition behind piecewise logistic regression is the same as that of asymmetric distributions, but is interpreted differently. In order to capture the characteristic that data points close to the boundary behave differently from data points at the two ends, in our experiments we use three-piece linear function in log-odds as the asymmetric Laplace method. However, the following formulation is for K pieces in general.

[1]It can be shown that θs which give maximum likelihood can only happen at individual data points.

The posterior probability of the model is defined as $P(y|x) = \frac{1}{1+\exp(-yf(x))}$ where $f(x)$ is a piecewise linear function of x. Following the idea from piecewise linear interpolation, we define $f(x)$ with $K+1$ knots ($\theta_0 < \theta_1 < \ldots < \theta_K$) as $f(x) = \sum_{j=0}^{K} w_j l_j(x)$, and $l_j(x)$ is defined as:

$$l_j(x) = \begin{cases} \frac{x-\theta_{j-1}}{\theta_j-\theta_{j-1}} & \theta_{j-1} \leq x < \theta_j \quad (j = 1, 2, \ldots, K) \\ \frac{x-\theta_{j+1}}{\theta_j-\theta_{j+1}} & \theta_j \leq x < \theta_{j+1} \quad (j = 0, 1, \ldots, K-1) \\ 0 & else \end{cases} .$$

It directly follows that our piecewise function $f(x)$ is continuous at all $K+1$ knots, with its value equals w_j at the j^{th} knot. We can use MLE to estimate the parameters (w_j's), which is equivalent to minimize the following objective $\mathbf{O}$:

$$\sum_{j=1}^{K} \sum_{x_i \in D_j} \log\{1+\exp(-y_i(w_j \frac{x_i-\theta_{j-1}}{\theta_j-\theta_{j-1}} + w_{j-1}\frac{x_i-\theta_j}{\theta_{j-1}-\theta_j}))\}$$

where $D_j = \{\theta_{j-1} \leq x < \theta_j\}$ ($j = 1, 2, \ldots, K$) is partition of D.

To avoid overfitting, we add a regularization term to the objective function

$$\mathbf{O}_{reg} = \mathbf{O} + \lambda \sum_{j=2}^{K} (\frac{w_j - w_{j-1}}{\theta_j - \theta_{j-1}} - \frac{w_{j-1} - w_{j-2}}{\theta_{j-1} - \theta_{j-2}})^2.$$

The term $\frac{w_j - w_{j-1}}{\theta_j - \theta_{j-1}}$ in the above formula is essentially the tangent of the linear piece over the subset D_j. Our intuition of adding the above penalization term is that tangents of adjacent linear pieces should not be too far away from each other, which allows the model to degenerate to standard logistic regression and avoids overfitting in local noisy regions. The parameter λ can be empirically chosen using a holdout dataset.

The remaining problem we have not addressed is how to choose the knots θ_j. It will be inefficient to do that as in asymmetric Laplace, due to the fact that for each candidate we need to train a model, rather than go through the data and compute the value directly. We can safely set $\theta_0 = \min\{x \in D\}$ and $\theta_K = \max\{x \in D\} + \epsilon$, where ϵ is a small positive number. Since our main experiment is to compare piecewise logistic regression with asymmetric Laplace, we now consider the case of three linear pieces with 4 knots: $\theta_0, \theta_1, \theta_2, \theta_3$ (cf. Section 3.4). We choose θ_1 from 10%, 20%, ..., 90% percentiles of the negative examples; and θ_2 is done over positive examples in the same manner. The optimal pair of knots is determined by maximum likelihood over training data, resulting in knots θ_1 and θ_2 that split the scores into the "hard to discriminate" decision area and two extremes.

To generalize the above idea to piecewise logistic regression with $K+1$ knots, we can first set the candidate knots using percentiles, then graduatelly add knot by maximizing likelihood with hill climbing. Since more knots will result in higher training likelihood (just as # of clusters in clustering), we need to do model selection to prevent overfitting.

3.4. Piecewise Logistic Regression vs Asymmetric Laplace

If we assume both positive and negative class-conditional densities are asymmetric Laplace with modes θ_1 and θ_2 respectively,

$$p(x|\theta_1, \beta_-, \gamma_-) = \begin{cases} \frac{\beta_-\gamma_-}{\beta_-+\gamma_-} \exp(-\beta_-(\theta_1 - x)) & x \le \theta_1 \\ \frac{\beta_-\gamma_-}{\beta_-+\gamma_-} \exp(-\gamma_-(x - \theta_1)) & x > \theta_1 \end{cases}$$

$$p(x|\theta_2, \beta_+, \gamma_+) = \begin{cases} \frac{\beta_+\gamma_+}{\beta_++\gamma_+} \exp(-\beta_+(\theta_2 - x)) & x \le \theta_2 \\ \frac{\beta_+\gamma_+}{\beta_++\gamma_+} \exp(-\gamma_+(x - \theta_2)) & x > \theta_2 \end{cases}$$

then the log-likelihood ratio (so does the log-odds) is a piecewise linear function with two knots at θ_1 and θ_2:

$$\log \frac{p(x|+)}{p(x|-)} = \begin{cases} c - \beta_+(\theta_2 - x) + \beta_-(\theta_1 - x) & x \le \theta_1 \\ c - \beta_+(\theta_2 - x) + \gamma_-(x - \theta_1) & \theta_1 < x \le \theta_2 \\ c - \gamma_+(x - \theta_2) + \gamma_-(x - \theta_1) & x > \theta_2 \end{cases}$$

where $c = \log \frac{\beta_+\gamma_+}{\beta_++\gamma_+} - \log \frac{\beta_-\gamma_-}{\beta_-+\gamma_-}$.

It is obvious that both our piecewise logistic regression (with 3 pieces, 4 knots) and the asymmetric Laplace distribution have the flexibility to model the log-odds with three pieces of linear functions rather than globally linear one as in standard logistic regression, and both of them have four parameters to estimate. However, they have quite different assumptions and use different criteria to optimize the model parameters. It is the main goal of this paper to compare these two methods in probability estimation.

3.5. Other Alternatives

It is possible to use richer family like higher order polynomials to model the log-odds $f(x)$, for example, Generalized Additive Model (Hastie & Tibshirani, 1996). Although this will give a better fit to training set, it does not necessarily improve the prediction performance due to the bias-variance trade-off. One nice property of piecewise logistic regression is that once the knots are fixed, the optimization is strict convex, which guarantees unique and global solution.

Another general way for probability estimation is to first estimate the empirical log-odds, then apply regression method to fit the resulting curve. However, obtaining a good estimation of log-odds is a difficult task, and we do not prefer this option by following Vapnik's principle (Vapnik, 1999) that "When solving a given problem one should avoid solving a more general problem as an intermediate step".

4. Experimental Design

To evaluate our proposed method, we compare it with the standard logistic regression and asymmetric Laplace method. We did not include other candidates like Gaussian, asymmetric Gaussian or Laplace since previous study (Bennett, 2003) showed that they are less effective compared with asymmetric Laplace and logistic regression.

4.1. Datasets

In order to compare our methods with previous ones (Bennett, 2003) on calibrating text classifiers' outputs, we use the same text categorization datasets (Reuters-21578 and TREC-AP) except for the one (MSN Web Directory collection) which is not publicly available.

4.1.1. Reuters-21578

Reuters-21578 has been one of the most widely used collections for text categorization, which contains new articles from 1987. We used the standard ModApte train/test split of 9603/3299 documents. There are altogether 135 categories, but only 90 of them appear in both training and test sets. In our experiments we used the ten most frequent categories, which allows us to compare our results to previous ones (Platt, 1999; Bennett, 2003).

4.1.2. TREC-AP

Our second data collection is the TREC-AP collection, which is a subset of the AP newswire stories of the TREC/TIPSTER data collection from 1988 through 1990. As in previous studies (Lewis et. al, 1996; Bennett, 2003), we use documents in 1988 and 1989 (142,791 documents) as our training set, and documents in 1990 (66,992 documents) as our test set. The categories are defined by keywords in a keyword field, and the title and body fields together are treated as documents. We use all twenty categories in this collection whose detailed information can be found in (Lewis et. al, 1996).

4.2. Classifiers

We use three classifiers, namely Naive Bayes, linear SVM and logistic regression to evaluate our score calibration methods. Linear SVM is one of the top performing classifiers that does not output probabilistic scores; although Naive Bayes can output posterior probabilities, due to its oversimplified assumptions, those probabilities are usually poor and skewed (Domingos & Pazzani, 1996). Different from the other two classifiers, logistic regression (Zhang & Oles, 2001) is a popular classification algorithm which can output well-calibrated probabilities. By adding it as one of our classifiers we are able to explore two interesting things: 1) Can we get further improvement by recalibrating well-calibrated scores generated by logistic regression classifier? 2) How well does the quality of probabilities of "classifier+calibration" compare with that of logistic regression classifier?

4.2.1. Naive Bayes

We use the rainbow package (McCallum, 1996) to perform the Naive Bayes classification, which is a multinomial model (McCallum & Nigam, 1998). Word probabilities are smoothed with Laplace smoothing (or equivalently, with a Dirichlet prior), and class priors are estimated from the training data distribution. We use the log-odds $\log \frac{P(+|\mathbf{d})}{P(-|\mathbf{d})}$ (Bennett, 2003) as the input score $x(\mathbf{d})$ of calibration methods.

4.2.2. Linear SVM

We use the SVM^{light} package (Joachims, 1998a) to train a linear kernel SVM model for the classification tasks. We use the default cost parameter C, which is taken to be $(\frac{1}{N}\sum_{i=1}^{N}||\mathbf{d}_i||)^{-2}$. The raw output of SVM $x(\mathbf{d}) = \sum_i \alpha_i K(\mathbf{d}, \mathbf{d}_i)$ is used as the input of calibration methods.

4.2.3. Logistic Regression

We use a variant of Gauss-Seidel (Zhang & Oles, 2001) to implement the logistic regression classifier, and the posterior probability $P(+|\mathbf{d}) = \frac{1}{1+\exp(-yf(\mathbf{d}))}$ is used as the input $x(\mathbf{d})$ of calibration methods.

4.3. Details of Experiments

In our experiments stopwords are removed, words are not stemmed, and rare words (happened less than three times in the corpus) are removed. We also select top 1000 words with Information Gain (Yang & Pederson, 1997) as our feature selection method. For linear SVM and logistic regression classifiers we use binary term weighting, and for Naive Bayes classifier no feature weighting is needed. To avoid overfitting, we use five-fold cross validation to obtain scores over the training set.

4.4. Evaluation Measures

We used three evaluation measures to examine these calibration methods, which are standard measures used in previous studies (Platt, 1999; Zadrozny & Elkan, 2001; Bennett, 2003) for evaluating the quality of probability estimations.

Our primary evaluation measure is **log-loss**, which is defined for each data instance x as:

$$L_{log-loss}(x) = \tilde{P}(+|x)\log P(+|x) + \tilde{P}(-|x)\log P(-|x)$$

where $\tilde{P}(+|x)$ and $\tilde{P}(-|x)$ are the empirical probabilities (they are either 0 or 1 based on the class label), and $P(.|x)$ is the model prediction. Notice that this is the only measure we want to maximize, though it is named "loss" as in the literature. The second evaluation measure is **squared-error**, which is defined as:

$$L_{squared-error}(x) = (\tilde{P}(+|x) - P(+|x))^2.$$

The last measure we used is **0/1-loss** (classification error), which gives the quantity of how calibrations can affect classification errors with respect to a fixed threshold 0.5 (except for SVM it is 0.0):

$$L_{0/1-loss} = \mathbf{I}((\tilde{P}(+|x) - 0.5)\cdot(P(+|x) - 0.5) > 0)$$

These three measures tend to have different properties. The **log-loss** will severely penalize extreme wrong decisions (for example, it is $+\infty$ when $\tilde{P}(+|\mathbf{x}) = 1$ and $P(+|\mathbf{x}) = 0$) over slightly wrong decisions. The **squared-error** uses squares of difference to measure how wrong the model prediction is from the empirical probability. Unlike the previous two measures, **0/1-loss** only cares how output probabilities behave locally around threshold 0.5, not measuring how far the wrong decisions are from the correct ones. By reporting results using all these measures, we can provide a thorough comparison of those methods.

5. Experimental Results

We conducted experiments for the calibration methods with classifiers (NB, SVM and LR) on Reuters-21578 and TREC-AP datasets. For each classifier, we report four sets of results, one for each of the calibration methods, plus the one called "raw" which is directly evaluated using the classifier's output without calibration. The log-loss and squared-error results for SVM raw output is not available, since SVM does not output

Table 1. Results for Reuters-21578: The best entry is in bold font, and a * means it is significantly better than other methods (using the same classifier).

classifier	method	log-loss	squared-error	0/1-loss
NB	raw	-32554.7	1586.8	1696
	LR	-2470.5	618.1	828
	ALaplace	-2140.2	575.8	788
	PLR	**-1964.1***	**556.5**	**780**
SVM	raw	N/A	N/A	492
	LR	-1577.3	396.3	511
	ALaplace	-1604.2	386.5	473
	PLR	**-1437.9***	**371.8**	**472**
LR	raw	-1882.6	462.3	612
	LR	-2819.1	642.0	832
	ALaplace	-inf	552.3	621
	PLR	**-1875.9**	**457.1**	**575***

Table 2. Results for TREC-AP

classifier	method	log-loss	squared-error	0/1-loss
NB	raw	-1193312.9	40379.4	45425
	LR	-33753.9	7788.6	9082
	ALaplace	-33391.4	7243.6	**8693**
	PLR	**-28315.1***	**6961.2***	8797
SVM	raw	N/A	N/A	7645
	LR	-27287.2	6036.9	7067
	ALaplace	-27487.9	**5724.2***	**6821***
	PLR	**-25972.5***	5916.7	7067
LR	raw	-32767.3	6682.2	**7705**
	LR	-35618.9	6935.1	**7705**
	ALaplace	-inf	8755.7	9722
	PLR	**-29153.9***	**6334.4***	7707

posterior probabilities. Our main results are shown in Table 1 & 2.

Previous studies (Bennett, 2003) have used significant sign test to evaluate the results. The way it worked is to look at wins and losses on decisions for two methods. However, applying the sign test for log-loss and squared-error to compare generative models and discriminative models is inappropriate due to the following reason: since both log-loss and squared-error are monotonically determined by the posterior probability, using them in the sign test is equivalent to using the posterior probabilities directly. After all, sign test ignores the magnitude of the estimation errors, it is too coarse a measure for evaluating how far the estimated probabilities are from the true probabilities. Therefore, we use the t-test to measure the significance for log-loss and squared-error, and use sign test to measure the significance only for 0/1-loss.

Our results in log-loss show that piecewise logistic regression is significantly better (significant level 0.05) than the other two methods under all conditions. In most cases it also performs the best in squared-error. Asymmetric Laplace in some cases achieved good performance in 0/1-loss, however, 0/1-loss should be treated lightly for the reasons discussed before.

In Figure 2 we plot empirical log-odds versus log-odds fittings generated by the calibration methods for categories **earn** in Reuters-21578 and **yugoslavia** in TREC-AP. We can see that piecewise logistic regression generates better fitting than the other two methods, which is generally the case for other categories as well. One interesting finding is that for all runs of piecewise logistic regression, the tangents of its three pieces are never negative (though theoretically speaking it might be), which is due to both the global monotonic trend and the regularization term that controls the tangent differences between adjacent pieces. This is generally believed to be a preferred property (Platt, 1999) for probability estimation. We also observed that for well-calibrated scores generated by LR classifier, only piecewise logistic regression further improved the performance (Table 1 & 2). The bad performance of asymmetric Laplace is because that the score distributions of LR classifier (ranged from 0 to 1) severely violates the assumption of asymmetric Laplace. Maybe the right way is to use the log-odds output of LR classifier to fit asymmetric Laplace, and then we probably end up with very similar result to that of SVM. This result also reflects that fact that discriminative approaches are generally more robust than generative ones.

Another advantage of the piecewise logistic regression is that it can have more than three linear pieces in the log-odds. We also did experiment to explore the effects of more than three pieces. As mentioned before, we first specify the candidate knots, which are the 5%, 10%, ..., 95% percentiles of D. Then knots (except the two end ones) are added one by one using hill-climbing strategy, by maximizing likelihood. To do model selection we can apply AIC, BIC or MDL like techniques, but here we use the following heuristic: if adding one more knot does not improve the likelihood by 1%, then the process is stoped and the previous model is selected as the final one. However, experiments showed that in most cases our algorithm ended with 2-4 pieces, and the results are very similar to what we get in Table 1 & 2 with 3 pieces. We think the reason why three-piece linear function usually suffices might be that the score distributions generated by those classifiers do not reflect very complicated shapes, as can be seen from Figure 2.

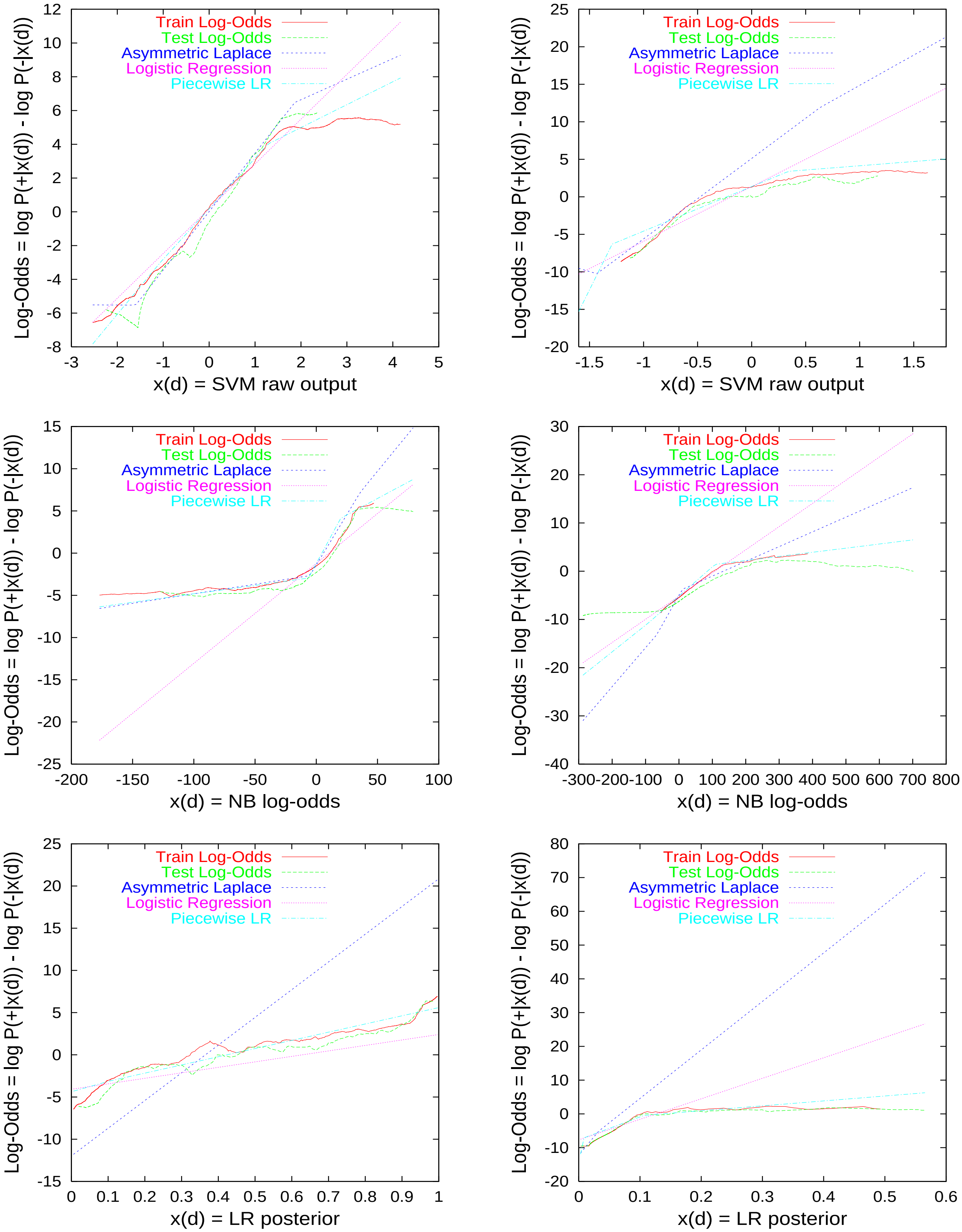

Figure 2. Empirical log-odds of the training/test data compared to the fitting generated by three methods. The left three graphs are for category **earn** in Reuters-21578, and the right three graphs are for category **yugoslavia** in TREC-AP. From top to bottom, graphs are for classifiers SVM, NB and Logistic Regression respectively.

6. Conclusions

In this paper we first reviewed two methods for score calibration, namely asymmetric Laplace and logistic regression, from generative and discriminative families respectively. Motivated by asymmetric Laplace which allows piecewise linear functions in log-odds, we proposed a novel and effective method, Piecewise Logistic Regression, as an alternative approach for probability estimation. By fitting a piecewise linear function in log-odds, this method has the same flexibility as asymmetric Laplace but makes fewer assumptions.

We provided an analysis showing the connections and differences between piecewise logistic regression and asymmetric Laplace, and also evaluated both methods as well as standard logistic regression with three classifiers on two standard text categorization collections. Our results showed that piecewise logistic regression performs significantly better than the other two calibration methods in log-loss metric across all the collections and classifiers tested. Moreover, the calibrated SVM probabilities achieved better results than those directly generated by logistic regression classifier.

Acknowledgements

We are grateful to Paul Bennett for providing code for generating log-likelihood scores from Naive Bayes and plotting empirical log-odds. We also thank anonymous reviewers for helpful comments. This work is sponsored in part by NSF under the grant KDI-9873009 and IIS-9982226. However, any opinions or conclusions in this paper are the authors' and do not necessarily reflect those of the sponsors.

References

Bennett, P. (2003). Using Asymmetric Distributions to Improve Text Classifier Probability Estimates. *Proceedings of SIGIR'03.*

Domingos, P., & Pazzani, M. (1996). Beyond independence: Conditions for the optimality of the simple bayesian classifier. *ICML'96.*

Duda, R., Hart, P., & Stork, D. (2001). Pattern Classification. John Wiley & Sons, Inc., 2001.

Hastie, T., & Tibshirani, R. (1996). Generalized Additive Model. *Statistical Sciences*, vol 1:297-318.

Hastie., T., & Tibshirani, R. (1996). Classification by pairwise coupling. *Technical Report*, Stanford University and University of Toronto.

Joachims, T. (1998a). Text categorization with support vector machines: Learning with many relevant features. *Proceedings of ECML.*

Kotz, S., Kozubowski T., & Podgorski K. (2001). *The Laplace Distribution and Generalizations: A Revisit with Applications to Communications, Economics, Engineering, and Finance.* Brikhauser.

Lewis, D., Schapire, R., Callan, J., & Papka, R. (1996). Training algorithms for linear text classifiers. *Proceedings of SIGIR'96..*

Lewis, D., & Gale, W. (1994). A sequential algorithm for training text classifiers. *SIGIR94.*

McCallum, A. (1996). Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. *http://www.cs.cmu.edu/~ mccallum/bow.*

McCallum, A., & Nigam, K. (1998). A comparison of event models for naive bayes text classification. *AAAI'98 Workshop on Learning for Text Categorization.*

Ng, A., & Jordan M. (2002). On Discriminative vs. Generative classifiers: A comparison of logistic regression and naive Bayes. *Proceedings of NIPS 14.*

Platt. J. (1999). Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. *Advances in Large Margin Classifiers*, MIT Press.

Rubinstein, Y., & Hastie, T. (1997). Discriminative vs. informative learning.p *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining* , pages 49-53. AAAI Press, 1997.

Vapnik, V. (1999). *The Nature of Statistical Learning Theory, 2nd edition.* Springer Verlag.

Yang, Y., & Pedersen, J. (1997). A Comparative Study of Feature Selection in Text Categorization. *ICML97.*

Zadrozny, B., & Elkan, C. (2001). Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. *Proceedings of the Eighteenth International Conference on Machine Learning*, 2001.

Zadrozny, B., & Elkan, C. (2002). Transforming Classifier Scores into Accurate Multiclass Probability Estimates. *SIGKDD*, 2002.

Zhang, T., & Oles, F. (2001). Text Categorization based on regularized linear classification methods. *Information Retrieval*, vol 4:5-31.