 © Physica-Verlag/Springer 2004

# THE TRADE-OFF BETWEEN GENERATIVE AND DISCRIMINATIVE CLASSIFIERS

**Guillaume Bouchard and Bill Triggs**

*Key words*: Statistical computing, numerical algorithms.

*COMPSTAT 2004 section*: Classification.

**Abstract**: Given any generative classifier based on an inexact density model, we can define a discriminative counterpart that reduces its asymptotic error rate. We introduce a family of classifiers that interpolate the two approaches, thus providing a new way to compare them and giving an estimation procedure whose classification performance is well balanced between the bias of generative classifiers and the variance of discriminative ones. We show that an intermediate trade-off between the two strategies is often preferable, both theoretically and in experiments on real data.

## 1 Introduction

In supervised classification, inputs $x$ and their labels $y$ arise from an unknown joint probability $p(x, y)$. If we can approximate $p(x, y)$ using a parametric family of models $\mathcal{G} = \{p_\theta(x, y), \theta \in \Theta\}$, then a natural classifier is obtained by first estimating the class-conditional densities, then classifying each new data point to the class with highest posterior probability. This approach is called *generative* classification.

However, if the overall goal is to find the classification rule with the smallest error rate, this depends only on the conditional density $p(y|x)$. *Discriminative* methods directly model the conditional distribution, without assuming anything about the input distribution $p(x)$. Well known generative-discriminative pairs include Linear Discriminant Analysis (LDA) vs. Linear logistic regression and naive Bayes vs. Generalized Additive Models (GAM). Many authors have already studied these models e.g. [5,6]. Under the assumption that the underlying distributions are Gaussian with equal covariances, it is known that LDA requires less data than its discriminative counterpart, linear logistic regression [3]. More generally, it is known that generative classifiers have a smaller variance than.

Conversely, the generative approach converges to the best model for the joint distribution $p(x, y)$ but the resulting conditional density is usually a biased classifier unless its $p_\theta(x)$ part is an accurate model for $p(x)$. In real world problems the assumed generative model is rarely exact, and asymptotically, a discriminative classifier should typically be preferred [9, 5]. The key argument is that the discriminative estimator converges to the conditional density that minimizes the negative log-likelihood classification loss against the true density $p(x, y)$ [2]. For finite sample sizes, there is a bias-variance tradeoff

COMPSTAT DRAFT

and it is less obvious how to choose between generative and discriminative classifiers.

In this paper, we will first consider the parameter estimation problem, focusing on the theoretical distinction between generative and discriminative classifiers. Then we propose a new technique for combining the two classifiers: the Generative-Discriminative Trade-off (GDT) estimate. It is based on a continuous class of cost functions that interpolate smoothly between the generative strategy and the discriminative one. Our method assumes a joint density based parametrization $p_\theta(x, y)$, but uses this to model the conditional density $p(x|y)$. The goal is to find the parameters that maximize classification performance on the underlying population, but we do this by defining a cost function that is intermediate between the joint and the conditional log-likelihoods and optimizing this on training and validation sets.

Given that the generative model based on maximum likelihood (ML) produces minimum variance — but possibly biased — parameter estimates, while the discriminative one gives the best asymptotic classification performance, there are good reasons for thinking that an intermediate method such as the GDT estimate should be preferred. We illustrate this on simulations and on real datasets.

## 2 Preliminaries

Using independent training samples $\{x_i, y_i\}, i = 1, \ldots, n$, $x_i \in \mathbb{R}^d$,and $y_i \in \{1, \ldots, K\}$ sampled from the unknown distribution $p(x, y)$, we aim to find the rule that gives the lowest error rate on new data. This is closely related to estimating the conditional probability $p(y|x)$.

For each of the $K$ classes, the class-conditional probability $p(x|y = k)$ is modeled by a parametric model $f_k$ with parameters $\theta_k$. The $y$ follows a multinomial distribution with parameters $p_1, \ldots, p_K$. The full parametrization of the joint density is $\theta = (p_1, \ldots, p_K, \theta_1, \ldots, \theta_K)$. Given $\theta$, new data points $x$ are classified to the group $k$ giving the highest posterior probability

$$P_\theta(Y = k|X = x) = \frac{p_k f_k(x; \theta_k)}{\sum_{l=1}^{K} p_l f_l(x_i; \theta_l)}. \tag{1}$$

The generative and the discriminative approaches differ only in the estimation of $\theta$.

**Generative classifier.** Given data $\{x_i, y_i\}$, $i = 1, \ldots, n$, a standard way to estimate the parameters of densities is the Maximum Likelihood (ML) estimate (we assume that the solution is unique):

$$\hat{\theta}_{GEN} = \arg\max_{\theta \in \Theta} \mathcal{L}_{GEN}(\theta), \qquad \mathcal{L}_{GEN}(\theta) = \textstyle\sum_{i=1}^{n} \log p_{y_i} f_{y_i}(x_i; \theta). \tag{2}$$

COMPUTER DRAFT

**Discriminative classifier.** Let $\mathcal{D} = \{p_\theta(y|x) = p_\theta(x,y)/\sum_z p_\theta(x,z), \theta \in \Theta\}$ be the set of conditional densities derived from the generative model. Our aim is to find the conditional density in $\mathcal{D}$ that minimizes a classification loss function on the training set. Here, we consider only the negative conditional log-likelihood $-\mathcal{L}_{DISC}$, which can be viewed as a convex approximation to the error rate:

$$\hat{\theta}_{DISC} = \arg\max_{\theta\in\Theta} \mathcal{L}_{DISC}(\theta), \qquad \mathcal{L}_{DISC}(\theta) = \sum_{i=1}^n \log \frac{p_{y_i} f_{y_i}(x_i;\theta)}{\sum_k p_k f_k(x_i;\theta)}. \quad (3)$$

The discriminative approach allows to eliminate parameters that influence only $p(x)$ ,not $p(y|x)$ (e.g. shared covariance matrix in Gaussian distributions), leading to logistic regression over lower dimensional parameter spaces. However, we will not use this reduction, as we need to maintain a common parametrization for the discriminative and generative cases. thus, the solution (3) of the discriminative classifier may not be unique — there may exist infinitely many parameters that give the same conditional distribution $p_\theta(x|y)$. However, the classification performance is the same for all such solutions.

**Relationship.** The quantity $\mathcal{L}_{DISC}$ can be expanded as follows:

$$\mathcal{L}_{DISC}(\theta) = \underbrace{\sum_{i=1}^n \log p_{y_i} f_{y_i}(x_i;\theta)}_{\mathcal{L}_{GEN}(\theta)} - \underbrace{\sum_{i=1}^n \log \sum_{k=1}^K p_k f_k(x_i;\theta)}_{\mathcal{L}_x(\theta)} \quad (4)$$

The difference between the generative and discriminative objective functions $\mathcal{L}_{GEN}$ and $\mathcal{L}_{DISC}$ is thus $\sum_{i=1}^n \sum_k \log p_\theta(x_i,k)$, the log-likelihood of the input space probability model $p_\theta(x)$. Equation (4) shows that compared to the discriminative approach, the generative strategy tends to favor parameters that give high likelihood on the training data.

## 3 Between Generative and Discriminative classifiers

To get a natural trade-off between the two approaches, we can introduce a new objective function $\mathcal{L}_\lambda$ based on a parameter $\lambda \in [0,1]$ that interpolates continuously between the discriminative and generative objective functions:

$$\begin{aligned} \mathcal{L}_\lambda(\theta;\boldsymbol{x},\boldsymbol{y}) &= \mathcal{L}_{GEN}(\theta;\boldsymbol{x},\boldsymbol{y}) - (1-\lambda)\mathcal{L}_x(\theta;\boldsymbol{x}) & (5)\\ &= \lambda\mathcal{L}_{GEN}(\theta) + (1-\lambda)\mathcal{L}_{DISC}(\theta). & (6)\end{aligned}$$

For $\lambda \in [0,1]$, the GDT estimate is

$$\hat{\theta}_\lambda = \arg\max_{\theta\in\Theta} \mathcal{L}_\lambda(\theta). \quad (7)$$

COMPANION DRAFT

Taking $\lambda = 0$ leads to the discriminative estimate $\hat{\theta}_{DISC}$, while $\lambda = 1$ leads to the generative one $\hat{\theta}_{GEN}$. We expect that the GDT estimates $\hat{\theta}_\lambda$ ($0 < \lambda < 1$) will sometimes have better generalization performances than these two extremes. Even if the discriminative estimate (3) is not unique, the maximum of (7) is unique for all $\lambda \in [0, 1)$ if the ML estimate $\hat{\theta}_{GEN}$ is unique.

**Computation of $\hat{\theta}_\lambda$.** Since we use a differentiable classification loss, the maximization problem (7) can be solved by any gradient ascent method. The Newton algorithm converges rapidly, but requires the computation of the Hessian matrix, The Conjugate Gradient (CG) algorithm may be more suitable for large scale problems: it needs only the first derivative and it is possible to avoid the storage of the quasi-Hessian matrix which can be huge when the number of parameters is large.

For simplicity, we assume that the parameters $\theta_k$ of the different class densities are independent. Taking the derivative of (5) with respect to $\theta_k$ and $\pi_k$, we get

$$\begin{cases} \frac{\partial}{\partial \theta_k} \mathcal{L}_\lambda(\theta_k) = \sum_{i=1}^n (\mathbf{I}_{\{y_i=k\}} - (1-\lambda)\tau_{ki}) \frac{\partial \log f_k(x_i;\theta_k)}{\partial \theta_k} \\ \frac{\partial}{\partial \pi_k} \mathcal{L}_\lambda(\theta_k) = \frac{1}{\pi_k} \left(n_k - (1-\lambda) \sum_{i=1}^n \tau_{ki}\right) \end{cases} \tag{8}$$

with $n_k = \sum_{i=1}^n \mathbf{I}_{\{y_i=k\}}$ and $\tau_{ki} = \frac{\pi_k f_k(x_i;\theta_k)}{\sum_{l=1}^K \pi_l f_l(x_i;\theta_l)}$. The optimal parameters are zeros of the equations (8) for $k = 1, \ldots, K$.

For a given class $k$, these equations are analogous to the ML equations on weighted data, although unlike ML, the weights can be negative here Each point has a weight $\mathbf{I}_{\{y_i=k\}} - (1-\lambda)\tau_{ki}$. The examples that have most influence on the $\theta_k$-gradient are those that belong to the class $k$ but have a low probability to be in it ($\tau_{ki}$ is small), and conversly those that do not belong to the class $k$ but that are assigned to it with a high probability. The influence of the assignment probabilities is controlled by the parameter $\lambda$. This remark may ultimately help us to link our approach to boosting, and similar algorithms that iteratively re-weight misclassified data. It also shows that the generative estimator ($\lambda$=1) is not affected by the classification rate of the data points.

**Choice of $\lambda$.** The GDT estimate contains a tuning parameter to set, which functions like the smoothing parameter in regularization methods. $\lambda$ cannot be set on the basis of minimum classification loss on the training set, since by definition, $\lambda = 0$ gives the optimal $\theta$ for training set classification. Instead, $\lambda$ is set to the value $\hat{\lambda}$ that minimizes the cross-validation error rate.

If the optimal $\hat{\lambda}$ is close to one, the generative classifier is preferred. This suggests that the bias in $p_\theta(x, y)$ (if any) does not affect the discrimination of the model too much. Similarly, if $\hat{\lambda}$ is close to 1, it suggests that the model $p_\theta(x, y)$ does not fit the data well, and the bias of the generative classifier is

CRC PRESS DRAFT

too high to provide good classification results. In this case, a more complex model — i.e. with more parameters, or less constrained — may be needed to reduce the bias. For other $\hat{\lambda}$, there is an equilibrium between the bias and the variance, meaning that the model complexity is well adapted to the amount of training data.

## 4 Simulations

To illustrate the behavior of the GDT method, we study its performance on two synthetic test problems. We define the true distributions of the data as follows: In the first experiment, the class conditional probabilities are gaussian with identity covariance matrix and means $m_1 = (1.25, 0, 0, 0)$ and $m_0 = (-1.25, 0, 0, 0)$. In the second case, we simulate $x$ according to a uniform density with correlated covariates : $x^{(1)} \sim \mathcal{U}[0; 1]$ and $x^{(d)} \sim \mathcal{U}[x^{(d-1)}; 1+x^{(d-1)}]$ with $d \in \{2, 3, 4\}$ and $x^{(i)}$ denotes the $i^{th}$ covariate. Then $y|x$ is simulated according to a Bernoulli distribution with parameter $1/\exp(-2.5x^{(1)})$. Note that the linear logistic model is true in the two experiments.

The assumed model is a Gaussian distribution for each class with shared diagonal covariance matrices and prior probabilities equal to $\frac{1}{K}$. Hence, the model does not correspond exactly to the true density in the second experiment, but it can provide a good approximation when the differences between the variances are small.

In each case, we estimated the true error rate of the classifiers learned on training samples of size 50, 100 and 200. The results are plotted in figure (1). We used standard plug-in estimates for $\lambda = 1$ and closed form logistic regression for $\lambda = 0$. For intermediate estimates, the conjugate gradient method was used. The first row illustrates the fact that the generative classifier performs better than the other estimates, but this difference tends to decrease when the sample size increases. In the second row, the best performance is from the BDG estimate for all training set sizes, and the optimal value of $\lambda$ (the one that minimizes the expected loss) decreases with $n$ since we know that the discriminative approach becomes optimal when $n$ tends to infinity.

## 5 Experiments

We tried our classification method on some of the publically available Statlog datasets. In our implementation of the GDT estimates, the parameter dimension is limited due to the size of the optimization problem (7). To make the computation feasible, we reduced the dimension of the data by computing the first four Fisher discriminant variables and using them as inputs (when the number of classes was less than 5, so that there were fewer than four discriminant directions, we computed the remaining directions by PCA on the residuals). These directions are computed using the training data and do not involve the test data.

We tried four types of density for the class-conditional distributions:

COMPUTER DRAFT

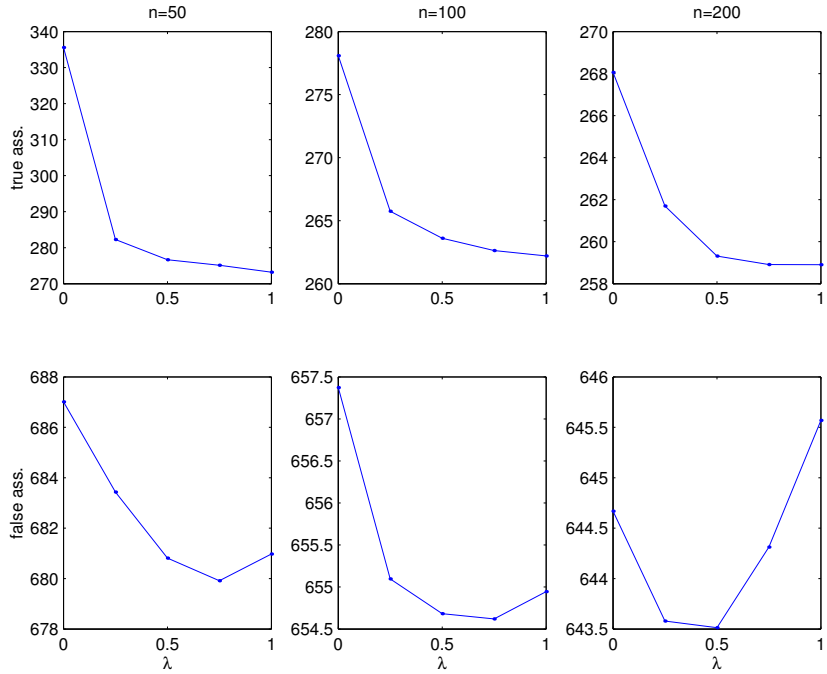


Figure 1: The full lines plot logistic loss computed on test sets of size $10^5$ against the tuning parameter $\lambda$. Each plotted value is the median of 200 experiments. The rows correspond to the first and second simultations. The columns correspond to different training sample sizes.

1. Gaussian densities with common covariance matrix (LDA), 2. Gaussian densities with unconstrained covariance (QDA), 3. Gaussian densities with spherical covariance matrix (Balls1), 4. Mixture of two Gaussian densities with spherical covariance matrix (Balls2). These distributions do not exactly fit the data, but they are distributions that are often used to approximate real datasets. Therefore, when the training sample is small, the generative approach may still behave better than the discriminative one. Training sample sizes were set to 50 times the number of classes so the discriminative classifiers should not have reached their asymptotic behavior.

We used a Cholesky-based parametrization of the inverse covariance matrix, so there was no need for a separate positivity constraint on the parameters. Derivatives with respect to this parametrization were obtained for each density, and we used the generative solution — which is explicit for densities 1-3 and obtained by the EM algorithm for the densities 4 — to initialize the CG algorithm.

Table (1) shows the generalization performance for each dataset and each model with different values of $\lambda$. These results show that substantial im-

DRAFT

COPY PAPER DRAFT

| Dataset | australian | diabetes | heart | satimage | vehicle |
|---|---|---|---|---|---|
| Training size | 100 | 100 | 100 | 300 | 200 |
| LDA GEN | **0.143** | 0.253 | 0.178 | 0.188 | 0.237 |
| LDA GDT0.75 | 0.144 | 0.252 | 0.178 | 0.187 | 0.235 |
| LDA GDT0.5 | 0.144 | **0.249** | 0.179 | 0.186 | 0.235 |
| LDA GDT0.25 | 0.144 | 0.250 | 0.182 | 0.185 | 0.236 |
| LDA DISC | 0.145 | 0.249 | 0.185 | 0.191 | 0.243 |
| QDA GEN | 0.149 | 0.262 | 0.181 | 0.181 | 0.235 |
| QDA GDT0.75 | 0.151 | 0.261 | 0.182 | **0.179** | 0.234 |
| QDA GDT0.5 | 0.150 | 0.262 | 0.181 | 0.180 | 0.235 |
| QDA GDT0.25 | 0.151 | 0.262 | 0.182 | 0.181 | 0.234 |
| QDA DISC | 0.168 | 0.270 | 0.204 | 0.215 | 0.267 |
| Balls1 GEN | 0.146 | 0.262 | 0.168 | 0.185 | 0.318 |
| Balls1 GDT0.75 | 0.145 | 0.260 | 0.167 | 0.183 | 0.293 |
| Balls1 GDT0.5 | 0.144 | 0.259 | **0.165** | 0.182 | 0.271 |
| Balls1 GDT0.25 | 0.144 | 0.257 | 0.169 | 0.181 | 0.254 |
| Balls1 DISC | 0.150 | 0.253 | 0.190 | 0.194 | 0.242 |
| Balls2 GEN | 0.146 | 0.266 | 0.181 | 0.185 | 0.239 |
| Balls2 GDT0.75 | 0.145 | 0.265 | 0.180 | 0.185 | 0.239 |
| Balls2 GDT0.5 | 0.146 | 0.265 | 0.180 | 0.184 | 0.236 |
| Balls2 GDT0.25 | 0.146 | 0.268 | 0.181 | 0.183 | **0.232** |
| Balls2 DISC | 0.166 | 0.279 | 0.211 | 0.210 | 0.250 |

Table 1: Test error rate on real datasets, averaged over 100 trials. For each trial, training data were randomly chosen and the error rate was computed on the remaining data. In the *heart* dataset, a misclassified heart disease has a cost of 5 instead of 1.

provements in the classification rate can be obtained for intermediate values of $\lambda$. However, they do not directly show the performance of the GDT estimate because we fixed $\lambda$ rather than selecting it by cross-validation on the training set. The evaluation of the cost as a function of $\lambda$ could be used as a model selection criterion. For example, on the vehicle dataset, the simple Gaussian model (Balls1) gives an optimal $\lambda$ equal to 0. This suggests that the bias is dominating the error, and indeed the results are improved by using two Gaussian densities for each class (Balls2).

One can object that the gain in error rate in these experiments is not sufficient to really conclude the usefulness of the GDT estimator.

## 6 Conclusion

In this study, the relationship between generative and discriminative classifiers has been clarified: they correspond to two different maximizations in the parameter space. By interpolating linearly between the two objective functions, we introduced the GDT estimator. This can be seen either as a less biased variant version of the discriminative solution, or as an improvement of the generative classifier. The regularization is "natural" in the sense that the parameters are encouraged to fit the inputs. Our preliminary results on real data showed that the intermediate model often gives better classification performances than the discriminative and generative classifiers.

The real interest of the GDT estimate resides in its application to generative models. Probabilistic models already exist in many areas: time series models, mixed models and graphical models — including Markov Random Fields and Hidden Markov Models — are examples of widely used generative models. When class-conditional probabilities are modelled generatively, then the GDT estimator should often improve the classification performances.

Currently, the main difficulty with the GDT method is the choice of the tuning parameter, as this requires an expensive cross-validation computation. We believe that more computationally efficient criteria can be developed by analyzing the solutions on the training set, in the spirit of the Bayesian Information Criterion [7].

## References

[1] Devroye L., Györfi L. & Lugosi L. (1997) *A* probabilistic Theory of Pattern Recognition. pp. 270-276 New York: Springer-Verlag.

[2] Efron. B. (1975) The efficiency of logistic regression compared to Normal Discriminant Analysis. *Journ. of the Amer. Statist. Assoc.*, 70:892-898.

[3] Ng A.Y. & Jordan M.I. (2002) On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes. In T. Dietterich, S. Becker and Z. Ghahramani (Eds.), *A*dvances in Neural Information Processing Systems 14, pp. 609-616. Cambridge, MA: MIT Press.

COMPUTER DRAFT

[4] Rubinstein Y.D. & Hastie T. (1997) Discriminative vs. informative learning. *I*n Proc. of the Third International Conference on Knowledge and Data Mining, pp. 49-53. AAAI Press.

[5] Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics 6,* pp. 461-464.

[6] Vapnick, V.N. (1998) *Statistical Learning Theory.* John Wiley & Sons.

*Acknowledgement*: We thank G. Celeux, for valuable discussions. This work was supported by the European LAVA project. *Address*: IS2 project
INRIA
38334 Saint-Ismier Cedex
`Guillaume.Bouchard@inria.fr`
*E-mail*: `Guillaume.Bouchard@inria.fr`

C04-PAPER DRAFT