Fisher's Conditionality Principle in Statistical Pattern Recognition

Carey E. PRIEBE

We present a simple, illustrative example of Fisher's Conditionality Principle in statistical pattern recognition. We observe training data $\{(X_i, Y_i, Z_i)\}_{i=1}^n$ with which to learn the discriminant boundary. At classification time, we observe the to-be-classified feature vector X with true-but-unobserved class label Y. We do not observe the Z associated with X, and the collection $\{Z_i\}$ is ancillary for the discriminant boundary. Nonetheless, $\{Z_i\}$ is essential for optimal classification.

KEY WORDS: Ancillary; Classification; Control variate.

INTRODUCTION

In a parametric model, an ancillary statistic is one whose distribution does not depend on the parameter of interest. While such statistics are sometimes mistakenly characterized as "useless" or "irrelevant," they are in fact neither (see Ghosh, Reid, and Fraser 2010 for a recent comprehensive review of ancillary statistics). Simple examples illustrating the utility of ancillary statistics for inference abound, but we know of no such example crafted explicitly in the context of statistical pattern recognition, in which the parameter of interest is the discriminant boundary. We present a simple, illustrative example of the utility of ancillary statistics in statistical pattern recognition wherein we observe training data $\{(X_i, Y_i, Z_i)\}_{i=1}^n$ and to-be-classified feature vector X with true-but-unobserved class label Y. We do not observe the Z associated with the to-be-classified X, and the collection $\{Z_i\}$ is ancillary for the discriminant boundary in our model; still, $\{Z_i\}$ is essential for optimal classification.

AMARI'S STATEMENT OF THE CONDITIONALITY PRINCIPLE

We consider Fisher's Conditionality Principle (Fisher 1950, 1956), and in particular Amari's statement of the Conditionality Principle (Amari 1985, p. 217): "When there exists an exact ancillary statistic r, the conditionality principle requires that the statistical inference should be performed by conditioning on r ...." The relevant point here is that aspects of the data which may seem to be not germane to the inferential task at hand are nonetheless valuable—essential, even. Amari continues (inference about u is the goal): "...A statistical problem then is decomposed into subproblems in each of which r is fixed at its observed value, thus dividing the whole set of the possible data points into subclasses. It is expected that each subclass consists of relatively homogeneous points with respect to the informativeness about u. We can then evaluate our conclusion about u based on r, and it gives a better evaluation than the overall average one. This is a way of utilizing information which ancillary r conditionally carries."

THE CONDITIONALITY PRINCIPLE IN STATISTICAL PATTERN RECOGNITION

Consider

$$(X, Y, Z), \{(X_i, Y_i, Z_i)\}_{i=1}^n \overset{\text{iid}}{\sim} F_{\mu_0} \in \mathcal{F} = \{F_\mu : \mu \in \mathbb{R}\},$$

where the distributions $F_\mu \in \mathcal{F}$ are specified via

$$Y \sim \text{Bernoulli}(1/2)$$

and

$$(X, Z)|Y = y \sim N\left(\begin{bmatrix} \mu + (-1)^y \\ 0 \end{bmatrix}, \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix}\right).$$

That is, the prior probabilities of class membership are $\pi_y = P[Y = y] = 1/2$ for $y = 0, 1$ and the class-conditional joint distributions of (X, Z) given $Y = y$ are bivariate normal. We see that $X|Y = y \sim N(\mu + (-1)^y, 2)$, $Z|Y = y \sim N(0, 1)$ (which implies that the unconditional $Z \sim N(0, 1)$, hence Z is ancillary), and $X|Z, Y = y \sim N(\mu + (-1)^y + Z, 1)$.

To apply the conditionality principle in statistical pattern recognition, the parameter of interest is the discriminant boundary. In this simple example, the discriminant boundary (when observing test datum X only and with the knowledge that the $\pi_y = 1/2$ and the class-conditional distributions of X are homoscedastic Gaussians) is a point $\mu \in \mathbb{R}$. The Bayes optimal classifier (upon observing X only) is given by

$$g^*(X) = I\{X < \mu_0\}$$

with Bayes optimal probability of misclassification given by

$$L^* = P[g^*(X) \neq Y] = \Phi(-1/\sqrt{2}).$$

It remains only to learn (estimate) the true but unknown μ_0 from the training data.

Since Z is ancillary for μ in $\mathcal{F}$ and is not observed for the test data, it might be naively thought that the training $\{Z_i\}$ can be safely discarded. That this is not so will be demonstrated here.

Carey E. Priebe is Professor, Department of Applied Mathematics and Statistics, Johns Hopkins University, Whitehead Hall, Baltimore, MD 21218-2682 (E-mail: *cep@jhu.edu*).

 DOI: 10.1198/tast.2011.09220

Discarding the ancillary $\{Z_i\}$ and relying on the distribution $X|Y$, the maximum likelihood estimate for the optimal discriminant boundary μ_0 is given by

$$\widehat{\mu} = (1/n)\sum_{i=1}^{n}(X_i - (-1)^{Y_i})$$

and using this estimate in the plug-in decision rule yields

$$\widehat{g}\big(X; \{(X_i, Y_i)\}_{i=1}^n\big) = I\{X < \widehat{\mu}\}.$$

Utilizing the ancillary $\{Z_i\}$ instead and relying on the distribution $X|Z, Y$, the maximum likelihood estimate for μ_0 is given by

$$\widetilde{\mu} = (1/n)\sum_{i=1}^{n}(X_i - (-1)^{Y_i} - Z_i)$$

which yields

$$\widetilde{g}\big(X; \{(X_i, Y_i, Z_i)\}_{i=1}^n\big) = I\{X < \widetilde{\mu}\}.$$

This latter estimate, $\widetilde{\mu}$, considers "the experiment actually performed" (utilizing the $\{Z_i\}$, here) by relying on $X|Z, Y$ rather than the marginal $X|Y$, in accordance with the conditionality principle.

Our measure of classifier performance is the conditional probability of misclassification error given training data $\mathcal{T}_n$ (Devroye, Gyorfi, and Lugosi 1997, p. 2),

$$L_n(g) = P[g(X; \mathcal{T}_n) \neq Y|\mathcal{T}_n].$$

Both $\widehat{\mu}$ and $\widetilde{\mu}$ are unbiased for and consistent estimators of μ_0;

$$\widehat{\mu} \sim N(\mu_0, 2/n)$$

and

$$\widetilde{\mu} \sim N(\mu_0, 1/n).$$

The superiority (smaller variance) of the estimate $\widetilde{\mu}$ of the discriminant boundary is due to the fact that Z is correlated with X and is an example of the well-known control variate variance reduction technique; in this case the control variate is ancillary and its use is demanded by the conditionality principle. This variance reduction results in superior classification performance: $L_n(\widetilde{g})$ is stochastically smaller than $L_n(\widehat{g})$. Indeed, a second-order Taylor series approximation demonstrates that utilizing the (ancillary) $\{Z_i\}_{i=1}^n$ approximately halves the excess error for this simple example. For any unbiased estimate μ_n,

$$\begin{aligned} L(\mu_n) &= P[I\{X < \mu_n\} \neq Y|\mathcal{T}_n] \\ &= h(\mu_n) \\ &= h(\mu_0) + h'(\mu_0)(\mu_n - \mu_0) \\ &\quad + \frac{1}{2}h''(\mu_0)(\mu_n - \mu_0)^2 + o_p\left(\frac{1}{n}\right) \end{aligned}$$

so

$$E[L(\mu_n)] = L^* + c\,\mathrm{Var}[\mu_n] + o\left(\frac{1}{n}\right),$$

where $c = \frac{1}{2}h''(\mu_0)$. Thus

$$\frac{E[L_n(\widetilde{g})] - L^*}{E[L_n(\widehat{g})] - L^*} \to \frac{1}{2}$$

as $n \to \infty$.

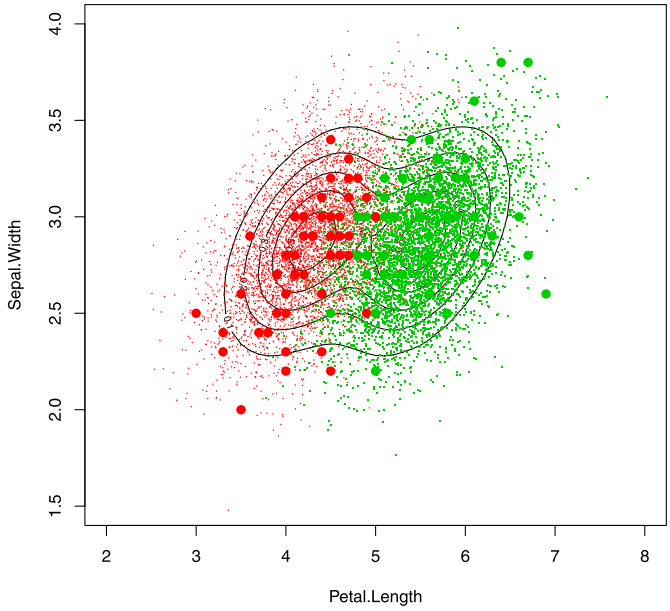

Figure 1. Synthetic example based on Fisher's Iris Data, with Petal Length playing the role of X and Sepal Width playing the role of the ancillary Z for the two classes Versicolor (depicted in red) and Virginica (depicted in green). The large dots are the 50 data points per class, and the small dots are generated from the synthetic distribution.

EXAMPLE

A simple example of our phenomenon in practice is given by a synthetic version of the well-known Fisher's Iris Dataset. Figure 1 plots the two features Petal Length (playing the role of X) and Sepal Width (playing the role of the ancillary Z) for the two classes Versicolor ($Y = 0$, depicted in red) and Virginica ($Y = 1$, depicted in green). The large dots are the 50 data points per class, and the small dots are random observations generated from the distribution

$$(X, Z)|Y = y \sim N\left(\begin{bmatrix} 4.906 + 0.646(-1)^y \\ 2.872 \end{bmatrix}, \begin{bmatrix} 0.260 & 0.076 \\ 0.076 & 0.100 \end{bmatrix}\right).$$

For this distribution, observing test observation X yields $L^* \approx 0.10$. Monte Carlo simulation with 10,000 replicates using $n = 10$ training observations $\{(X_i, Y_i, Z_i)\}_{i=1}^{10}$ results in

$$\frac{E[L_{10}(\widetilde{g})] - L^*}{E[L_{10}(\widehat{g})] - L^*} \approx 0.80;$$

the sign test indicates superiority of $\widetilde{g}$ with a p-value $< 10^{-10}$.

CONCLUSION

The purpose of this short note is to illustrate Fisher's Conditionality Principle and the utility of ancillary information in statistical pattern recognition using the simplest possible example. Generalization of the phenomenon to more realistic and applicable settings (multivariate X, more complex model/discriminant boundary) is straightforward (see, e.g.,

Priebe, Marchette, and Healy 2004 for such a setting) but the fundamental idea is here presented in its fullest simplicity.

[Received October 2009. Revised June 2011.]

REFERENCES

Amari, S.-I. (1985), *Differential Geometric Methods in Statistics. Lecture Notes in Statistics*, Vol. 28, Berlin: Springer. [167]

Devroye, L., Gyorfi, L., and Lugosi, G. (1997), *A Probabilistic Theory of Pattern Recognition*, New York: Springer. [168]

Fisher, R. A. (1950), *Contributions to Mathematical Statistics*, New York: Wiley. [167]

——— (1956), *Statistical Methods and Scientific Inference*, Edinburgh: Oliver and Boyd. [167]

Ghosh, M., Reid, N., and Fraser, D. A. S. (2010), "Ancillary Statistics: A Review," *Statistica Sinica*, 20, 1309–1332. [167]

Priebe, C. E., Marchette, D. J., and Healy, D. M. (2004), "Integrated Sensing and Processing Decision Trees," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26, 699–708. [169]