Binary Classification with Models and Data Density Distribution

Xuan Chen

Supervised by
Professor Raymond Chi-Wing Wong

Abstract

Traditionally, the study of binary classification has been formulated as a deterministic problem with 0-1 labels. However, probabilistic labels are becoming more popular nowadays, and they have many practical applications in our real life. Since probabilistic labels are more informative - they imply probabilities of the samples belonging to positive cases (i.e., labelled as 1), intuitively predictions can be more accurate with probabilistic labels in training datasets. Therefore, in this project, from both theoretical and experimental view **we verify whether probabilistic labels are worth using in different data density distribution.**

Problem Definition

Training datasets

Consider a binary classification with two classes, 0 and 1. In traditional setting, we are given a training dataset T which contains instances $I_1, I_2 \ldots I_n$. Each instance is associated with a feature vector $\boldsymbol{x_i}$ and a target attribute y_i where $i \in \mathbb{N}$. Let X be the set of all possible feature vectors. Note that there are two possible values of target attribute y_i, 0 and 1. A classifier h is defined to be a hypothesis which takes feature vector x_i as an input and its output y_i is either 0 or 1.

Everything of datasets with probabilistic labels remains the same, except that now target attributes are fractional scores $f_{i..}$ We assume that all instances are generated according to a joint distribution of two random variables, X and Y, denoted by $Pr(X, Y)$. Given a feature vector $\boldsymbol{x}$, we define $\eta(\boldsymbol{x})$ **to be the conditional probability $Pr(Y=1 \mid X=x)$, the probability that an instance with its feature having its target attribute equal to1**. Note that $\hat{\eta}(\boldsymbol{x})$ is the estimated probability, i.e., $\hat{\eta}(\boldsymbol{x})$ **can be considered as the computed version** of $\eta(\boldsymbol{x})$. **f_i can be regarded as an "observed" version of $\eta(\boldsymbol{x_i})$**, since it is obtained by labelers and statistical information. To be more specific, f_i is the value $\eta(\boldsymbol{x_i})$ added by Gaussian white noise. With this noise condition, each fractional score f_i follows $N(\eta(\boldsymbol{x_i}), \sigma^2)$. If f_i is smaller than 0, then it can be assigned to class 0. Likewise, if f_i is larger than 1, then it can be assigned to class 1.

Measurement of Error

Given **a classifier $\boldsymbol{h} = \boldsymbol{I}_{\hat{\eta}(x) \geq 0.5}$**, the expected error of h, denoted by $\mathrm{err}(h)$, is defined to be $Pr_{(x,y)\sim Pr(X,Y)}(y \neq h(\boldsymbol{x}))$. The Bayes classifier, **denoted by $\boldsymbol{h^*} = \boldsymbol{I}_{\eta(x) \geq 0.5}$**, is defined to be the classifier which gives the minimum expected error. Given a classifier, **its excess error, denoted by $E(h)$, is defined as $\mathrm{err}(h)$-$\mathrm{err}(h^*)$.** Note that $E(h)$ must be greater than 0. The hypothesis is more accurate when $E(h)$ is approaching 0.

Data Density Distribution

Firstly we here state the definition of Tysbakov Noise Condition:

Definition: Given two noise parameters $c>0$ and $\gamma \geq 0$, $\forall t \in (0, 0.5)$,

$$\Pr(\mathbb{E}[|\eta(x) - 0.5|] < t) \leq ct^{\gamma}$$

Define $f(t) = ct^{\gamma}$. $f'(t)$ reflects the distribution of data density.

1) Convex Distribution:
   a) Bowl-shape Distribution: when $\gamma > 2$, data density in terms of $\eta(\boldsymbol{x})$ looks like a bowl.
   b) V-shape Distribution: when $\gamma = 2$, data density in terms of $\eta(\boldsymbol{x})$ is a symmetrical V-shape and the lowest point is at $\eta(\boldsymbol{x})$=0.5.
   c) Casp Distribution: when $1 < \gamma < 2$, data density in terms of $\eta(\boldsymbol{x})$ is a casp.
2) Uniform Distribution: when $\gamma = 1$, data is uniformly distributed.
3) Peak Distribution: when $0 \leq \gamma < 1$, data accumulates around classification boundary.
4) Double-arch and Double-shape Distribution: in this case we write the definition formula as $\Pr(\mathbb{E}[|\eta(x) - 0.5|] < t) \leq c(\frac{1}{3}t^3 + \frac{1}{4}t^2)$ instead of original formula in order to get tighter error bound.

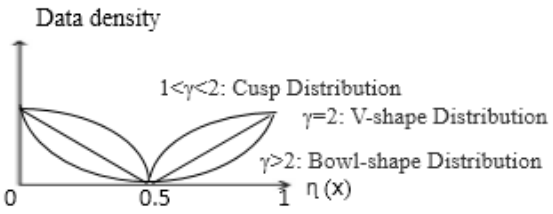


Figure 1: Convex Distribution

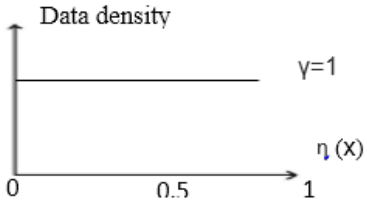


Figure 2: Uniform Distribution

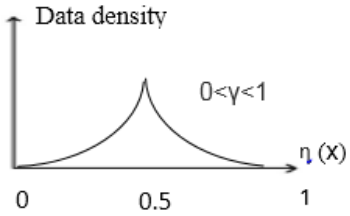


Figure 3: Peak Distribution

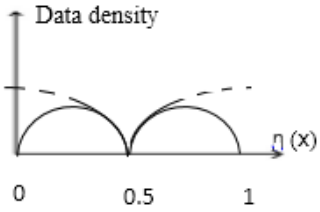


Figure 4: Double-arch Distribution

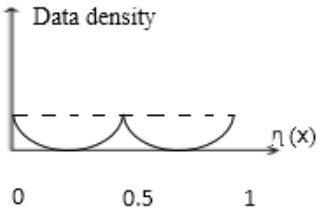


Figure 5: Double-bowl Distribution

Models

In this project, we theoretically analyze the error bound of prediction with the use of probabilistic labels and **Gaussian Process Regression**. We also adopted **Radial Basis Function Network, Nearest Neighbor** as well as **LibSVM** to our experiments for comparison.

Problem

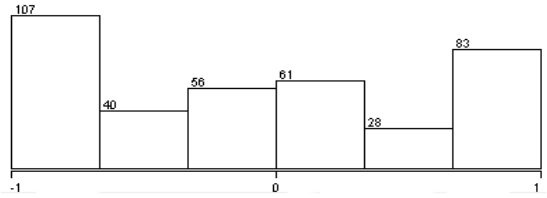
In this project, we study whether in every data density distribution can the models perform better with probabilistic labels than with clear-cut labels.

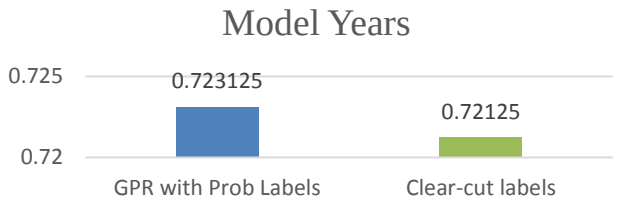
Theoretical Results

	Bowl-shape	V-shape	Cusp	Uniform	Peak	Double-arch	Double-bowl
Error bound found in this project	$< \tilde{O}(n^{-1})$	$= \tilde{O}(n^{-1})$	$> \tilde{O}(n^{-1})$ and $< \tilde{O}\left(n^{-\frac{3}{4}}\right)$	$\tilde{O}\left(n^{-\frac{3}{4}}\right)$	$> \tilde{O}\left(n^{-\frac{3}{4}}\right)$ and $\tilde{O}\left(n^{-\frac{1}{2}}\right)$	$= \tilde{O}(n^{-1})$	$= \tilde{O}(n^{-1})$
Best-known error bound under realizable setting					$= \tilde{O}(n^{-1})$		
Best-known error bound under non-realizable setting					$= \tilde{O}(n^{-\frac{1}{2}})$		

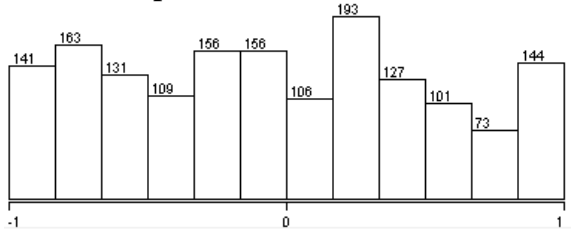
Experiments

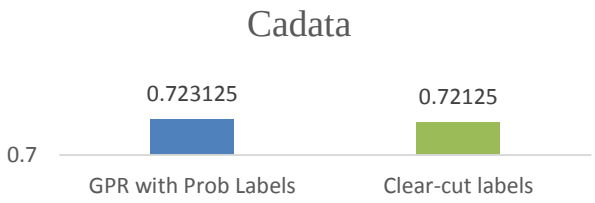
We conducted experiments via Weka 3.6, on a workstation with 2.10 GHz CPU and 2.0 GB RAM. To examine the prediction accuracy, we performed a 10-fold cross validation for these experiments. In the result we present the accuracy of GPR with probabilistic labels, and the best result with clear-cut labels. Datasets were derived from UCI repository.



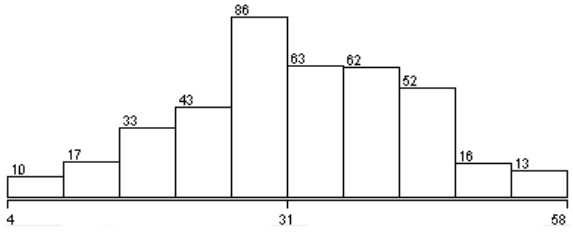


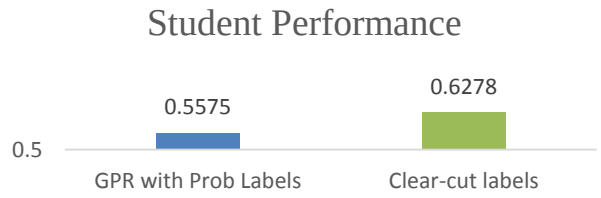
Data in *Model Years* roughly followed bowl-shape distribution. With probabilistic labels, the prediction was much more accurate than with clear-cut labels.



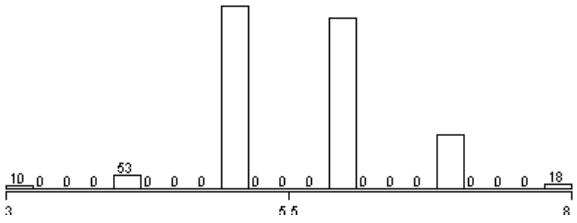


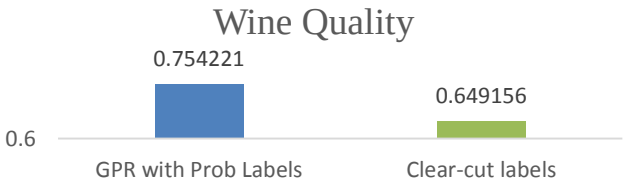
Data in *Cadata* roughly followed uniform distribution. With probabilistic labels, the prediction was no better than with clear-cut labels.





Data in *Student Performance* roughly followed peak-distribution. The accuracy was even high when clear-cut labels were involved in dataset.





Data in *Wine Quality* roughly followed double-arch distribution. With probabilistic labels, the prediction was apparently much more accurate than with clear-cut labels.

Conclusion

Only when data density follows any one of bowl-shape, V-shape, double-arch or double-bowl distribution do we get a prediction with Gaussian Process Regression and probabilistic labels at least as precise as with clear-cut labels. In other cases, the conclusion does not necessarily hold.