

# HOMEWORK 3

CMPS242 - FALL 2015

Ehsan Amid

Andrei Ignat

Sabina Tomkins

October 23, 2015

## 1 Problem One

To determine the error rate we look at the probability of a point being misclassified. There are two cases where a point can be misclassified i.e. we either label a + point with the - label or we label a - with the - point with the + label. Thus total error probability is equal to

$$P(\text{error}) = P(x = -) \cdot P(\text{label} = + | x = -) + P(x = +) \cdot P(\text{label} = - | x = +)$$

Since each point is drawn independently, the conditional probability of the label given the classification of the target point, is just the probability of the label. So we have:

$$P(x = +) \cdot P(\text{label} = -) + P(x = -) \cdot P(\text{label} = +)$$

where  $P(\text{label})$  is determined by the probability of the configurations given in Table ?? . We notice that there are four possible neighbor configurations:

Table 1: Configurations

Configurations	- - +	+ + -	- - -	+ + +
Probability of points in configuration	$\frac{1}{3} \cdot \frac{1}{3} \cdot \frac{2}{3}$	$\frac{1}{3} \cdot \frac{2}{3} \cdot \frac{2}{3}$	$\frac{1}{3} \cdot \frac{1}{3} \cdot \frac{1}{3}$	$\frac{2}{3} \cdot \frac{2}{3} \cdot \frac{2}{3}$
Number of ways to choose each configuration	$\binom{3}{2}$	$\binom{3}{2}$	1	1
Probability of each configuration	$\frac{6}{27}$	$\frac{12}{27}$	$\frac{1}{27}$	$\frac{8}{27}$

From Table ?? we can infer that the probabilities of each label that the algorithm predicts i.e.

$$P(\text{label} = -) = P(- - +) + P(- - -) = \frac{6}{27} + \frac{1}{27} = \frac{7}{27}$$

$$P(\text{label} = +) = P(- + +) + P(+ + +) = \frac{8}{27} + \frac{12}{27} = \frac{20}{27}$$

We plug these values into the formula we obtained for the total error probability and we obtain:

$$P(\text{error}) = P(x = -) \cdot P(\text{label} = +) + P(x = +) \cdot P(\text{label} = -)$$

$$= \frac{1}{3} \cdot \frac{20}{27} + \frac{2}{3} \cdot \frac{7}{27} = 0.4197 > \frac{1}{3}$$

This error is higher than the Bayes optimal, which is comforting as we know we cannot get lower than the optimal  $\frac{1}{3}$ . Without getting into the same level of detail, we note that, if the noise rate is  $\frac{1}{10}$ , this implies that the Bayes optimal hypothesis becomes  $\frac{1}{10}$  and, following the same argument as before, we get:

$$P(\text{error}) = P(x = -) \cdot P(\text{label} = +) + P(x = +) \cdot P(\text{label} = -)$$

$$= \frac{1}{10} \left( \binom{3}{2} \cdot \frac{1}{10} \cdot \frac{9}{10} \cdot \frac{9}{10} + \frac{9}{10} \cdot \frac{9}{10} \cdot \frac{9}{10} \right) + \frac{9}{10} \left( \binom{3}{2} \cdot \frac{1}{10} \cdot \frac{1}{10} \cdot \frac{9}{10} + \frac{1}{10} \cdot \frac{1}{10} \cdot \frac{1}{10} \right)$$

$$= 10^{-4} (3 \cdot 81 + 729 + 3 \cdot 81 + 9) = \frac{1224}{10000} = 0.1224 > \frac{1}{10}$$

We see that once again, the probability of error is higher than the Bayes optimal hypothesis. This is a comforting result.

## 2 Problem Two

## 3 Problem Three

We are sorry to ruin someone's hopes and dreams, but this is not a big achievement for the algorithm. A simple model would always predict the majority class. In this case, if the model always selected the majority class, you would have accuracy at least  $\frac{1000}{1010} = .990099$ . So for any model to be said to perform well, it should get at least 99% accuracy. A model with exactly 99% accuracy is doing nothing better than always predicting the majority class. So it will always misclassify the 10 points of the second class and will never be able to achieve 100%. We would like a model which can correctly classify the 10 points of the second class, as well as the 1000 points in the first class. The data is very much skewed, in this case, it is easy to achieve very high accuracy. So for skewed training data, accuracy alone is not a good metrics, precision and recall should also be measured.

## 4 Problem Four

We start by addressing the questions:

- If AP courses are taken, predict **H** if the GPA is between ..., and ...
- if AP courses are not taken, predict **H** if the GPA is between ..., and ...

We will predict **H**, if the probability of **H** given the data is greater than the probability of **N**. Thus we need to be able to evaluate the  $P(\mathbf{H}|\text{data})$  and the  $P(\mathbf{N}|\text{data})$ . Using Naive Bayes we can define  $P_H$  as the probability that the AP was taken, conditioned on the students being honors students. Similarly,  $P_N$  is the probability that the AP was taken conditioned on the students not being honors students.

$$\begin{aligned} P(\mathbf{H}|\text{GPA} = x, \text{AP} = \text{yes}) &= P(\text{GPA} = x|\mathbf{H}) \cdot P(\text{AP} = \text{yes}|\mathbf{H})P(\mathbf{H}) \\ &= \frac{1}{\sigma_H \sqrt{2\pi}} e^{-(x-\mu_H)^2/2\sigma_H^2} \cdot P_H^{\text{yes}} \cdot (1 - P_H)^{1-\text{yes}} \cdot P(\mathbf{H}) \\ P(\mathbf{N}|\text{GPA} = x, \text{AP} = \text{yes}) &= P(\text{GPA} = x|\mathbf{N}) \cdot P(\text{AP} = \text{yes}|\mathbf{N})P(\mathbf{N}) \\ &= \frac{1}{\sigma_N \sqrt{2\pi}} e^{-(x-\mu_N)^2/2\sigma_N^2} \cdot P_N^{\text{yes}} \cdot (1 - P_N)^{1-\text{yes}} \cdot P(\mathbf{N}) \end{aligned}$$

Our goal will be to compare these two equations before we plug in values for the  $\mu$ 's and  $\sigma$ 's. In order to not overcomplicate ourselves with incomprehensible math, we define  $k_H^{\text{yes}}$  and  $k_N^{\text{yes}}$  as

$$k_H^{\text{yes}} = \frac{P_H^{\text{yes}}(1 - P_H)^{1-\text{yes}}P(\mathbf{H})}{\sigma_H} \quad k_N^{\text{yes}} = \frac{P_N^{\text{yes}}(1 - P_N)^{1-\text{yes}}P(\mathbf{N})}{\sigma_N}$$

Our probabilities thus become:

$$\begin{aligned} P(\mathbf{H}|\text{GPA} = x, \text{AP} = \text{yes}) &= k_H^{\text{yes}} \frac{1}{\sqrt{2\pi}} e^{-(x-\mu_H)^2/2\sigma_H^2} \\ P(\mathbf{N}|\text{GPA} = x, \text{AP} = \text{yes}) &= k_N^{\text{yes}} \frac{1}{\sqrt{2\pi}} e^{-(x-\mu_N)^2/2\sigma_N^2} \end{aligned}$$

We now attempt to determine the GPA range based on the fact that, if the student is an **H**onors student, the following should hold

$$P(\mathbf{H}|\text{GPA} = x, \text{AP} = \text{yes})P(\mathbf{N}|\text{GPA} = x, \text{AP} = \text{yes}) \\ k_H^{\text{yes}} \frac{1}{\sqrt{2\pi}} e^{-(x-\mu_H)^2/2\sigma_H^2} > k_N^{\text{yes}} \frac{1}{\sqrt{2\pi}} e^{-(x-\mu_N)^2/2\sigma_N^2}$$

We simplify the  $\sqrt{2\pi}$ , and rearrange such that we get all the exponentials on one side and everything else on the other side. What we get is

$$\frac{e^{-(x-\mu_H)^2/2\sigma_H^2}}{e^{-(x-\mu_N)^2/2\sigma_N^2}} > \frac{k_N}{k_H}$$

The next natural step is to take the natural log(which also leads to a sign inversion since on both sides we have values smaller than 1) and this leads to:

$$\begin{aligned}
& -(x - \mu_H)^2 / 2\sigma_H^2 + (x - \mu_N)^2 / 2\sigma_N^2 < \ln \left( \frac{k_N}{k_H} \right) \\
& \frac{1}{2\sigma_H^2} \cdot (-x^2 + 2\mu_H x - \mu_H^2) + \frac{1}{2\sigma_N^2} \cdot (x^2 - 2\mu_N x - \mu_N^2) + \left( \frac{\mu_N^2}{2\sigma_N^2} - \frac{\mu_H^2}{2\sigma_H^2} \right) < \ln \left( \frac{k_N}{k_H} \right) \\
& x^2 \left( \frac{1}{2\sigma_N^2} - \frac{1}{2\sigma_H^2} \right) + x \left( \frac{\mu_H}{\sigma_H^2} - \frac{\mu_N}{\sigma_N^2} \right) + \left( \frac{\mu_N^2}{2\sigma_N^2} - \frac{\mu_H^2}{2\sigma_H^2} - \ln \left( \frac{k_N}{k_H} \right) \right) < 0
\end{aligned}$$

Thus we have to solve a quadratic equation of the form  $ax^2 + bx + c = 0$  with

$$\begin{aligned}
a &= \frac{1}{2\sigma_N^2} - \frac{1}{2\sigma_H^2} \\
b &= \frac{\mu_H}{\sigma_H^2} - \frac{\mu_N}{\sigma_N^2} \\
c &= \left( \frac{\mu_N^2}{2\sigma_N^2} - \frac{\mu_H^2}{2\sigma_H^2} - \ln \left( \frac{k_N}{k_H} \right) \right) = \left( \frac{\mu_N^2}{2\sigma_N^2} - \frac{\mu_H^2}{2\sigma_H^2} - \ln \left( \frac{P_N^{\text{yes}}(1 - P_N)^{1-\text{yes}} P(\mathbf{N}) \sigma_H}{P_H^{\text{yes}}(1 - P_H)^{1-\text{yes}} P(\mathbf{H}) \sigma_N} \right) \right)
\end{aligned}$$

To find the values for a, b, and c, we consider the data in Table ??.

Table 2: Class and GPAs

Class	AP	GPA
H	yes %	4.0 %
H	yes %	3.7 %
H	no %	2.5 %
N	no %	3.8 %
N	yes %	3.3 %
N	yes %	3.0 %
N	no %	3.0 %
N	no %	2.7 %
N	no %	2.2 %

We now solve for the mean and variance using Maximum Likelihood and formulas 2.121 and 2.122 from the Bishop textbook and get the results from Table ??

Table 3: Class and GPAs

Class	Mean	Variance
H	3.4	.4200
N	3.0	.2433

Similarly we can draw from the Bernoulli distribution which is defined as  $P(x) = P^x(1 - P)^{1-x}$ . In our case,  $P$  is actually  $P_H$  and  $P_N$  and thus these two Bernoulli distributions are based on the probabilities in the following table:

Table 4: Class and AP

Class	P
H	$P_H = 2/3$
N	$P_N = 1/3$

In order to compute our complicated logarithm term, give the value 1 to the case of the student having taken AP and 0 to the student not having taken AP. The only thing that is missing now is the general probability of having a honors/non-honors student which from Table ?? are  $P(\mathbf{H}) = \frac{1}{3}$  and  $P(\mathbf{N}) = \frac{2}{3}$ . We now have all the ingredients needed to obtain our coefficients  $a$ ,  $b$  and  $c$ . However, we now make the following observation and namely that all the formulas that we painstakingly derived so far stay the same when we consider the case of the student not having taken an AP class with the only difference being in the logarithm term. The logarithm term changes in the sense that, when the student is not an AP student, the  $(1 - P_H)$  and  $(1 - P_N)$  terms will affect the formula instead of the  $P_H$  and  $P_N$  terms which are "activated" when the student is an AP student. This leads us to the following values for  $a$ ,  $b$ , and  $c$ :

Table 5: Class and GPAs for AP=yes

AP	$a$	$b$	$c$
yes	5.6099	-31.3917	42.6870
no	5.6099	-31.3917	41.3007

Plugging these values into a numerical solver for the roots of the second order equation yields the following predictions to our initial question:

- If AP courses are taken, predict **H** if the GPA is between 2.3300, and 3.2658
- if AP courses are not taken, predict **H** if the GPA is between 2.1152, and 3.4806

A sanity check for a case in which the GPA would be 3.0 shows that for both the case of AP, as well as for non-AP, the second order equation is smaller than 0, given the coefficients  $a$ ,  $b$  and  $c$ .