# Stat 406 Spring 2007: Homework 2

Out Mon 15 Jan, back Mon 22 Jan

## 1 Gaussian decision Boundaries

Suppose we have two 1D normal distributions with the same variance, but with different means: $N(\mu_1, \sigma^2)$ and $N(\mu_2, \sigma^2)$. Explain the effect on the decision boundary of changing the class prior $p(Y = 1)$.

## 2 More Gaussian decision boundaries

Let $p(x|y = j) = \mathcal{N}(x|\mu_j, \sigma_j)$ where $j = 1, 2$ and $\mu_1 = 0, \sigma_1^2 = 1, \mu_2 = 1, \sigma_2^2 = 10^6$. Let the class priors be equal, $p(y = 1) = p(y = 2) = 0.5$.

1. Find the decision region
$$R_1 = \{x : p(x|\mu_1, \sigma_1) \geq p(x|\mu_2, \sigma_2)\} \tag{1}$$
Sketch the result. Hint: draw the curves and find where they intersect. Find *both* solutions of the equation
$$p(x|\mu_1, \sigma_1) = p(x|\mu_2, \sigma_2) \tag{2}$$
Hint: recall that to solve a quadratic equation $ax^2 + bx + c = 0$, we use
$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} \tag{3}$$

2. Now suppose $\sigma_2 = 1$ (and all other parameters remain the same). What is $R_1$ in this case?

## 3 Bayes classifier for Gaussian data

Note: you can solve this exercise by hand or using a computer (matlab, R, whatever). In either case, show your work. Consider the following training set of heights $x$ (in inches) and gender $y$ (male/female) of some US college students.

| x | y |
|---|---|
| 67 | m |
| 79 | m |
| 71 | m |
| 68 | f |
| 67 | f |
| 60 | f |

1. Fit a Bayes classifier to this data, using maximum likelihood estimation, i.e., estimate the parameters of the class conditional likelihoods
$$p(x|y = c) = \mathcal{N}(x; \mu_c, \sigma_c) \tag{4}$$
and the class prior
$$p(y = c) = \pi_c \tag{5}$$
What are your values of $\mu_c$, $\sigma_c$, $\pi_c$ for $c = m, f$? Show your work (so you can get partial credit if you make an arithmetic error).

2. Compute $p(y = m|x, \hat{\theta})$, where $x = 72$, and $\hat{\theta}$ are the MLE parameters. Hint: recall that a Gaussian density is given by

$$\mathcal{N}(x|\mu, \sigma) \stackrel{\text{def}}{=} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \tag{6}$$

3. What would be a simple way to extend this technique if you had multiple attributes per person, such as height and weight? Write down your proposed model as an equation.

## 4 Gaussian classifier for height/weight data

In this example, you will train a Bayesian classifier to compute $p(y|x)$, where $y \in \{1, 2\}$ representing male or female, and $x$ is either the person's height, weight, or both. When we combine height and weight, we will compare a full covariance model with a diagonal covariance model (naive Bayes). Turn in a printout (hardcopy) of your code and figures/ results.
Be sure to download `Code.zip`, and `Data.zip`. Suppose you unzip them to `C:/foo/Code` and `C:/foo/Data`. Then in matlab you can type `addpath C:/foo/Code` and `addpath C:/foo/Data`.

1. Load the data using `heightWeightDataLoad`. This returns `data.X`, where column one is height and column two is weight, and `data.Y`, where 1=male, 2=female. Partition this data into a training set (80% of the data) and a testing set (20%). You can use the provided function `partitionDataset` for this. To ensure everyone gets the same results, please set the random number seed as follows:

```
data = heightWeightDataLoad;
seed = 0;
rand('state', seed);
randn('state', seed);
[traindata, testdata] = partitionDataset(data, 0.8);
```

2. Using the provided function `gaussianClassifierTrain`, train up 4 different classifiers on the training data. Model 1 uses the height, model 2 uses the weight, model 3 uses both, and model 4 is a naive Bayes classifier that uses both. Hint: model 4 can be derived from model 3 by making the covariance matrices be diagonal. (The covariance matrices are 3D matrices where `params.Sigma(:,:,c)` represents $\Sigma_c$ for class $c$.)

3. Using the provided function `gaussianClassifierApply`, apply your 4 models to the test set (using the appropriate columns of the test data) and compute $p_{im} = p(y = 1|x(i, :), m)$ for model $m$ and test case $i$. Now plot $p_{im}$ vs $i$ for each model, superimposing the plots. To make the results easier to interpret, first sort the test data so the males come before the females using

```
[junk, perm] = sort(testdata.Y); % 1's (male) come first
testdata.X = testdata.X(perm,:);
testdata.Y = testdata.Y(perm);
```

   The result should look like Figure 1. Everything to the left of index 20 is male and should be a large number; everything to the right of index 20 is female and should be a small number.

4. It is hard to tell which classifier is working best, so use the provided function `ROCcurve` to plot ROC curves for the 4 models. For example, for model $m$, you can use

```
[faRate{m}, hitRate{m}, AUC] = ROCcurve(probMale(:,m), (testdata.Y==1), 0);
```

   where `probMale(i,m)` = $p_{im}$ defined above. Plot hitRate vs faRate using `plot` and `hold on`. The result should look something like Figure 2.

5. The ROC curves show that weight is a better feature than height. However, earlier we showed that the $d'$ value for height was larger than for weight. Explain this discrepancy. Also, the naive Bayes result seems to be the same as the full covariance. Explain this discrepancy.

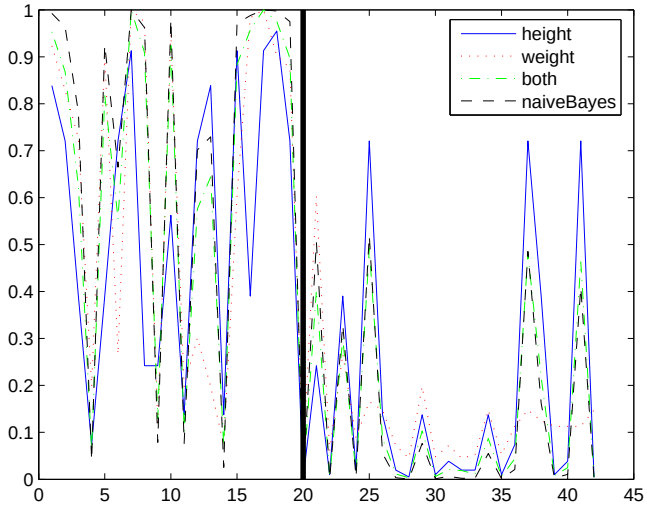


*Figure 1:* Probability of being male vs index on the test set.

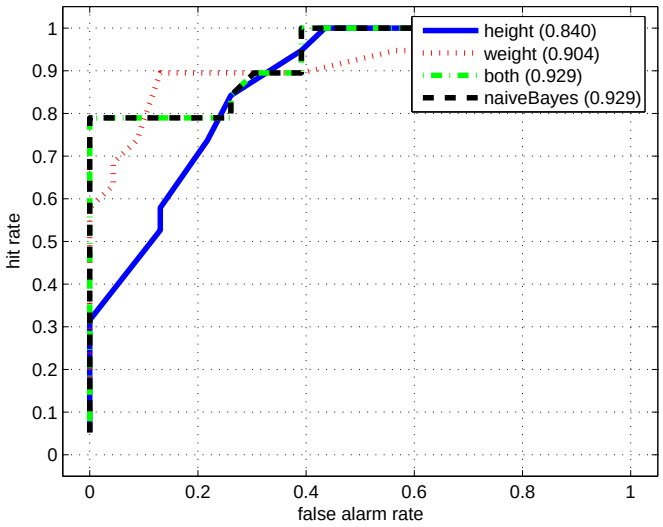


*Figure 2:* ROC curves for models 1 to 4.