DATA 1010
In-class exercises
Samuel S. Watson
09 November 2018

Problem 1

I did some consulting once for a businessman who told me that the way to distinguish a machine learning expert from a novice is to give them some of your data and see whether they can write a classifier which has very low error when applied to a withheld test set. Why was he wrong?

Solution

There is no guarantee that the features contain enough information for a low-test-error classifier to exist. In other words, even the Bayes classifier might have a high error rate.

In the context of the previous problem, for example, it is clear that no one can reliably predict a person's gender based on their height. My cousin is 5'9"—is he a man or a woman?

Problem 2

Consider a binary classification problem where the two classes are equally probable, the class-0 conditional density is a standard multivariable normal distribution in two dimensions, and the class-1 conditional density is a multivariate normal distribution with mean $[1, 1]$ and covariance I.

Find the class boundary for the Bayes classifier.

Solution

The Bayes classifier is $(x, y) \mapsto \text{argmax}_i \, p_i f_i(x, y)$, where

$$f_0(x, y) = \frac{1}{2\pi} e^{-\frac{1}{2}(x^2+y^2)}, \text{ and}$$
$$f_1(x, y) = \frac{1}{2\pi} e^{-\frac{1}{2}((x-1)^2+(y-1)^2)}.$$

By symmetry, the classifier will predict class 1 for every point above line $x + y = 1$. We can obtain the same result by setting $f_1 = f_2$ and solving:

$$-\frac{1}{2}(x^2 + y^2) = -\frac{1}{2}((x-1)^2 + (y-1)^2).$$

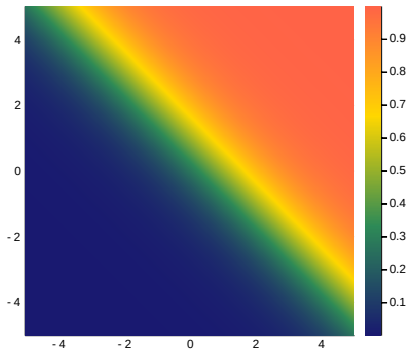
This equation simplifies to $x + y = 1$, as desired.

Problem 3

(Problem 2 continued). Find the regression function $r(x) = \mathbb{E}[Y \mid X = x] = \mathbb{P}(Y = 1 \mid X = x)$. Plot a heatmap of this function.

Solution

```
using Plots, Distributions, Optim
gr(aspect_ratio=1,color=cgrad([:MidnightBlue,:SeaGreen,:Gold,:Tomato]))
A = MvNormal([0,0],[1.0 0; 0 1])
B = MvNormal([1,1],[1.0 0; 0 1])
xs = -5:1/2^5:5
ys = -5:1/2^5:5
r(x,y) = pdf(B,[x,y])/(pdf(A,[x,y])+pdf(B,[x,y]))
rs = [r(x,y) for x=xs,y=ys]
heatmap(xs,ys,rs)
```



Problem 4

(Problem 2 continued). Sample 1000 points by choosing one of the two distributions uniformly at random and then sampling from the selected distribution. Model the regression function by fitting the parametric model

$$r(\mathbf{x}) = \sigma(\alpha + \boldsymbol{\beta} \cdot \mathbf{x}),$$

where $\sigma(x) = \frac{1}{1+e^{-x}}$. Use the loss functional

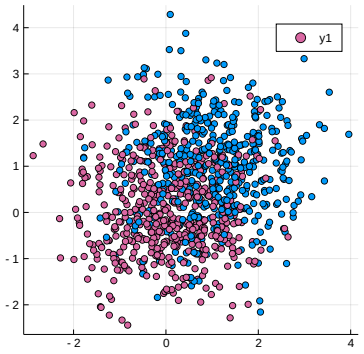
$$L(r) = \sum_{i=1}^{n} \left[y_i \log \frac{1}{r(x_i)} + (1 - y_i) \log \frac{1}{1 - r(x_i)} \right].$$

This is called **logistic regression**.

Solution

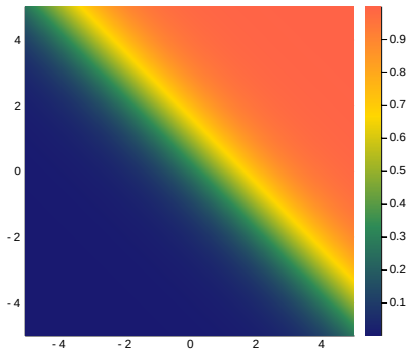
We begin by sampling the points (recall that the Unicode subscript i may be obtained by doing backslash-underscore-i-tab).

```
samples = [rand() > 1/2 ? (rand(A),0) : (rand(B),1) for i=1:1000]
xᵢs = [x for ((x,y),c) in samples]
yᵢs = [y for ((x,y),c) in samples]
cs = [c for ((x,y),c) in samples]
scatter(xᵢs,yᵢs,group=cs,color=cs)
```



Next, we define the loss function and minimize it:

```
σ(u) = 1/(1 + exp(-u))
r(β,x) = σ(β⋅[1;x])
C(β,xᵢ,yᵢ) = yᵢ*log(1/r(β,xᵢ))+(1-yᵢ)*log(1/(1-r(β,xᵢ)))
L(β) = sum(C(β,xᵢ,yᵢ) for (xᵢ,yᵢ) in samples)
β̂ = optimize(L,ones(3),BFGS()).minimizer
r̂s = [r(β̂,[x,y]) for x=xs,y=ys]
heatmap(xs,ys,r̂s)
```



Problem 5

(Problem 2 continued). Was the supposition that r took the given parametric form correct?

Solution

Yes! We can calculate

$$\frac{\mathrm{e}^{-\frac{x^2}{2}-\frac{y^2}{2}}}{\mathrm{e}^{-\frac{x^2}{2}-\frac{y^2}{2}}+\mathrm{e}^{-\frac{(x-1)^2}{2}-\frac{(y-1)^2}{2}}}=\frac{1}{1+e^{x+y-1}},$$

which does take the form $\sigma(\alpha+\boldsymbol{\beta}\cdot\mathbf{x})$.

Problem 6

How could we have modified the setup of Problem 2 so that our parametric assumption did not hold? Hint: begin by showing that the decision boundary is necessarily linear for a logistic model.

Solution

As suggested in the problem statement, we note that the set of points satisfying $\sigma(\alpha + \boldsymbol{\beta} \cdot \mathbf{x}) = \frac{1}{2}$ is the set of points where $\alpha + \boldsymbol{\beta} \cdot \mathbf{x} = -1$, which is a hyperplane.

If the covariance matrices of the two classes were different, then the boundary would have quadratic rather than linear. Therefore, the logistic regression would not have identified the correct boundary.

Problem 7

How could we modify the logistic regression model so that it would accommodate multivariate normal class conditional distributions with distinct covariance matrices?

Solution

We could supplement the features x_1 and x_2 with the three quadratic terms x_1^2, x_2^2, and $x_1 x_2^2$. The set of linear equations in these five quantities is equal to the set of quadratic equations in the original variables x_1 and x_2.