MACHINE LEARNING IN PHYSICS
FOUNDATIONS 4

HARRISON B. PROSPER

PHY6938

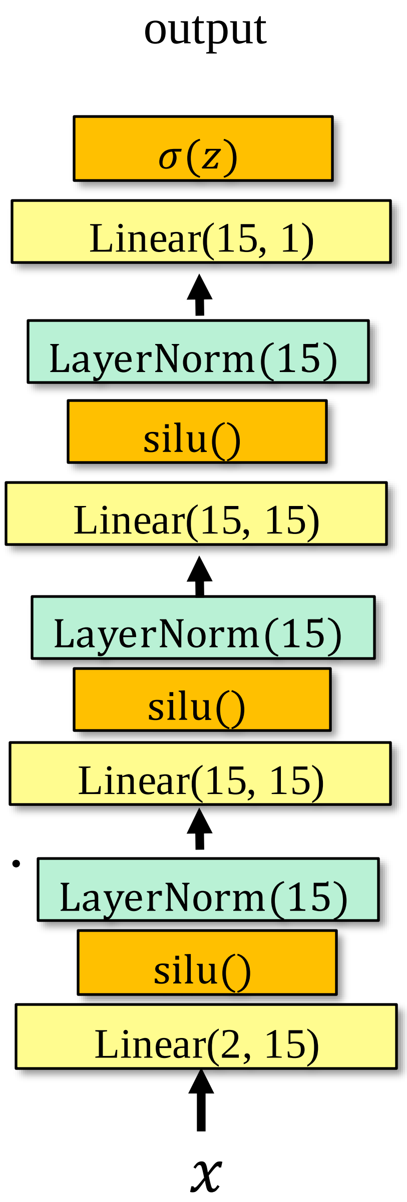
Recap

In Lab01, we trained the model on the right by minimizing the empirical risk

$$R(\omega) = \frac{1}{N}\sum_{i=1}^{N} L(y_i, f_i)$$

using the binary cross entropy loss

$$L(y, f) = -[y \log f + (1 - y)\log(1 - f)].$$



Recap

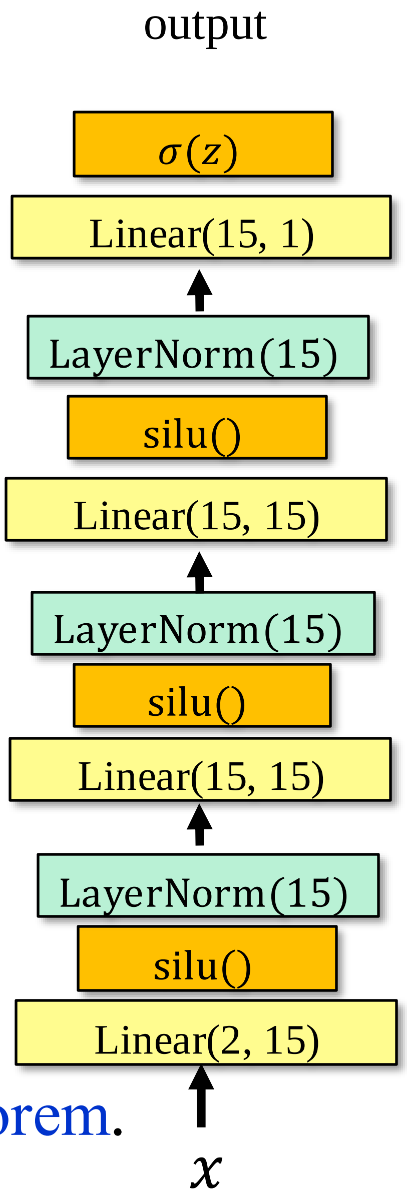
Using the general expression

$$\int dy \, \frac{\partial L}{\partial f} p(y \mid x) = 0$$

we found that *any* model $f(x, \omega)$ trained with the binary cross entropy *necessarily* approximates the probability $p(y = 1 \mid x)$

$$= \frac{p(x|y = 1) \, \pi(y = 1)}{p(x|y = 1)\pi(y = 1) + p(x|y = 0)\pi(y = 0)}$$

The expression above is an instance of Bayes' theorem.



BINARY CLASSIFIERS

Binary Classifiers

Let's start by simplifying our notation. Defining $p(k|x) \equiv p(y = k|x)$ and $\pi(k) \equiv \pi(y = k)$, we can write

$$p(y = 1 \mid x)$$

$$= \frac{p(x|y = 1)\, \pi(y = 1)}{p(x|y = 1)\pi(y = 1) + p(x|y = 0)\pi(y = 0)}$$

as

$$p(1 \mid x) = \frac{p(x|1)\, \pi(1)}{p(x|1)\pi(1) + p(x|0)\pi(0)}$$

Given a threshold t, the classification rule is

$$\text{class assignment} = \ 1 \text{ if } p(1 \mid x) > t \text{ else } 0.$$

Binary Classifiers

Let us pause for a moment to ask the following question.

In what sense is the rule

$$\text{class assignment} = \ 1 \text{ if } p(1 \mid x) > t \text{ else } 0$$

optimal?

Binary Classifiers

Assignment 1

For a 1-dimensional variable x, and a threshold x_0 corresponding to t, write an expression for the probability of misclassification (that is, the *error rate*) for a dataset with prior probabilities $\pi(1)$ and $\pi(0)$ and densities $p(x \mid 1)$ and $p(x \mid 0)$.

Minimize the error rate with respect to the threshold x_0 and show that this yields the rule $p(1 \mid x) > t$, called the Bayes' classifier.

Binary Classifiers

In Lab01, we used a dataset comprising two classes of object labeled either $y = 1$ or $y = 0$. Therefore, necessarily,

$$\pi(0) + \pi(1) = 1$$

Likewise, the probability $p(0 \mid x) \equiv p(y = 0|x)$, that is, an object characterized by the features x is of the class labeled $y = 0$, necessarily satisfies

$$p(0 \mid x) + p(1 \mid x) = 1.$$

Therefore,

$$p(0 \mid x) = \frac{p(x|0)\,\pi(0)}{p(x|1)\pi(1) + p(x|0)\pi(0)}$$

$p(1 \mid x)$ and $p(0 \mid x)$ are referred to as a class probabilities.

Binary Classifiers

In Lab01, used a balanced dataset, that is, a dataset with $\pi(1) = \pi(0) = \frac{1}{2}$.

For a balanced dataset, the probability $p(1 \mid x)$ simplifies to

$$D(x) \equiv \frac{p(x|1)}{p(x|1) + p(x|0)}$$

In physics, the function $D(x)$ is often referred to as a discriminant.

Binary Classifiers

Like the class probability, an object can be classified using the rule

$$\text{class assignment} = 1 \text{ if } D(x) > t \text{ else } 0.$$

A typical physics use case is classifying objects as signal or background, e.g., Type 1a supernovae versus the rest.

But is this rule optimal when the dataset has a large imbalance: $\pi(1) \ll \pi(0)$?

The answer is “yes”! Let’s see why…

Binary Classifiers

When $\pi(1) \neq \pi(0)$ the correct class probability is

$$p(1 \mid x) = \frac{p(x|1)\, \pi(1)}{p(x|1)\pi(1) + p(x|0)\pi(0)}$$

Divide the numerator and denominator by $p(x|1) + p(x|0)$.

This yields

$$p(1 \mid x) = \frac{D(x)\, \pi(1)}{D(x)\pi(1) + (1 - D(x))\pi(0)}$$

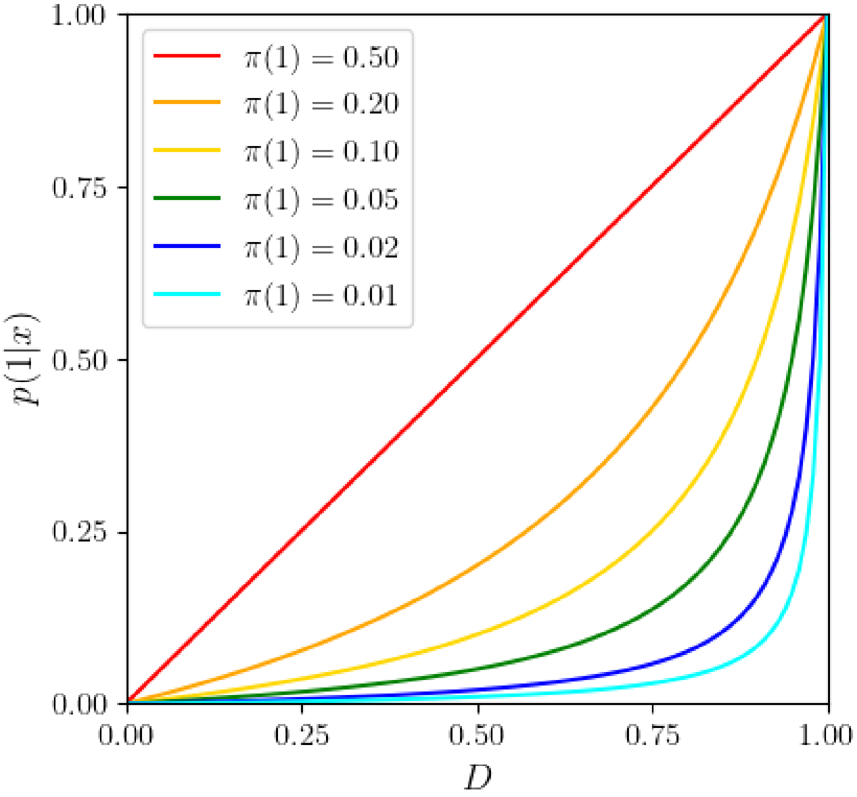
The key point to note is that $p(1 \mid x)$ is a *monotonic* function of $D(x)$. Therefore, $D(x)$ and $p(1 \mid x)$ classify equally well!

Binary Classifiers

$$p(1\,|\,x) = \frac{D(x)\,\pi(1)}{D(x)\pi(1) + (1 - D(x))(1 - \pi(1))}$$

The figure shows the dependence of $p(1\,|\,x)$ on $D(x)$ for different values of the prior $\pi(1)$.

Note: when $\pi(1) = \frac{1}{2}$ we get $p(1\,|\,x) = D(x)$ and, therefore, a straight line of unit slope.



HOW GOOD IS OUR CLASSIFIER?

How Good Is Our Classifier?

1. Machine learning models typically give only point estimates. The absence of a measure of confidence in their estimates is a serious deficiency, a point we shall take up towards the end of the semester.

2. The other problem is that at present we rely on heuristics to judge whether a training schedule has yielded a model that is as close as it can get to the optimal solution.

3. The third problem is that it is very difficult to define what constitutes a fair performance comparison between models.

How Good Is Our Classifier?

1. In contemporary ML applications, model performance is typically valued a lot more than model simplicity.

2. Suppose that a model with 10^6 parameters does 30% better, in some measure, than one with 5,000 parameters. If the 30% improvement is judged to be worthwhile the tendency is to favor the larger model.

3. For classifiers, the two most common measures of performance are:
    1. The Receiver Operating Characteristic (ROC) curve
    2. And the Area Under the Curve (AUC).

Summary

- A classifier minimizes the error rate.

- When created using a balanced dataset, we typically refer to the classifier as a discriminant.

- Classification using a discriminant is (in principle) just as powerful as classification using the correct class probability