On a surprising relationship between statistical physics and machine learning

Oct 1, 2021

Introduction

On August 24th, 2021 a very interesting paper was published in the Proceedings of the National Academy of Sciences with the title The Fermi–Dirac distribution provides a calibrated probabilistic output for binary classifiers [1]. The authors report on a "suprising relationship between the probability of a sample belonging to one of the two classes and the Fermi-Dirac distribution determining the probability that a fermion occupies a given single-particle quantum state in a physical system of noninteracting fermions." They describe a method for calibrating the probabilities predicted by a binary classifier and a method for ensembling disparate classifiers, each based upon this suprising relationship. The connection between binary classification and statistical physics is indeed remarkable but it is not novel. The authors are in fact discussing applications of the venerable method known as logistic regression. The authors claim that their method is distinct from logistic regression but they are mistaken. Let's trace through the paper with a view to understanding this surprising connection. The authors of the paper may be surprised to find that the end result is not that their research topic is less rich than they believe, but that logistic regression is a richer and more interesting method than they've imagined.

The sections on statistical physics will likely appear to dwell too much on the basics for those already familiar with the subject and likewise for the sections on statisical machine learning. This was unfortunately difficult to avoid because my goal is that this post be accessible to people from both audiences. Feel free to skim or skip through any passages that appear tedious to you.

Review of Fermi-Dirac Statistics

First let's give a brief introduction to the basics of Fermi-Dirac statistics. I'll try to make the presentation accessible to those from a general mathematically literate audience who may not have much experience with physics. We're only going through the minimum background necessary,

describing the mathematical properties of this model without going into any of the underlying physics.

The problem concerns systems of many indistinguishable and non-interacting elementary particles called fermions. Systems of non-interacting particles are also known as gases and the systems of interest are known as Fermi gases.

Each fermion can be thought of as having an energy associated to it which comes from a discrete set of nonzero values

$$\mathcal{E} = \{\epsilon_1, \epsilon_2, \ldots, \epsilon_m\}$$

without loss of generality we suppose

$$0 < \epsilon_1 < \epsilon_2 < \ldots < \epsilon_m$$

each fermion has a state associated to it which also comes from a discrete set of values

$$\mathcal{S} = \{s_1, s_2, \ldots, s_M\}$$

and there is mapping from states onto energy levels such that any particle in state s_i will have energy ϵ_j for some fixed j, but multiple states can have the same associated energy.

Suppose there are g_i states associated to energy level ϵ_i. g_i is referred to as the degeneracy of the system at energy level ϵ_i. The unique properties of a Fermi gas are due to fermions obeying something called the Pauli exclusion principle. This principle asserts that within a system, at most one particle can occupy any given state s_i. There is elegant math and physics lurking behind this principle but it is outside the scope of this post. We will take it as given within the mathematical model.

Fermi-Dirac statistics concerns the distribution of energy values for particles within a Fermi gas with known possible energy levels and where the total number of particles N and the total energy E within the system are known. N is assumed to be sufficiently large to allow for some simplifications that we will see shortly. The problem is to determine the distribution $\pi(\epsilon \,|N, E)$ describing the probability that a state s is occupied given its associated energy level ϵ and with fixed N and E.

There are multiple ways to derive the correct distribution. The most relevant for this post is based on an approach initiated by Austrian physicist Ludwig Boltzmann and developed by American physicist Josiah Willard Gibbs in the 19th century [2] and first applied to the problem of Fermi-Dirac statistics by the British physicist Paul Dirac in 1926 [3]. The correct distribution was also derived independently by the Italian physicist Enrico Fermi in 1926 by a different method [4].

Suppose that any arrangement of N total fermions among the $M = \sum_{i=1}^{m} g_i$ states in $\mathcal{S}$ is equally likely. Given a random distribution of N fermions among the M states, we ask:

Given a sequence of natural numbers

$$\mathbf{n} = (n_1, n_2, \ldots, n_m)$$

such that

$$\sum_{i=1}^{m} n_i = N$$

and

$$0 <= n_i <= g_i$$

for each i.

what is the probability $\pi(\mathbf{n})$ that n_i of the g_i states at energy level ϵ_i are occupied for each i?

The problem is in essence a simple combinatorial puzzle. The number of ways to distribute N identical fermions among M states is given by the binomial coefficient

$$\binom{M}{N} = \binom{g_1 + g_2 + \cdots + g_m}{N}$$

(recall that by the Pauli exclusion principle, at most one fermion can occupy any given state). Similarly, the number of ways to distribute n_i fermions among the g_i states at energy level ϵ_i is given by the binomial coefficient $\binom{g_i}{n_i}$. Through the rule of product we see that the total number of ways to distribute the N fermions among the M states with n_i fermions occupying the g_i states at energy level ϵ_i for $\mathbf{n}$ satisfying the properties above is

$$\mathrm{W}\left(\mathbf{n}\right) = \prod_{i=1}^{m} \binom{g_i}{n_i}$$

and thus

$$\pi(\mathbf{n}) = \frac{\mathrm{W}\left(\mathbf{n}\right)}{\binom{M}{N}}$$

The standard way to solve the problem from here is to attempt to find the value of $\mathbf{n}$ that maximizes the probability $\pi(\mathbf{n})$ given the additional constraint that the total energy is equal to E. You may have noticed that this is a challenging discrete problem and that it is ill posed in the sense that for many values of E and N there is no choice of $\mathbf{n}$ that even satisfies the constraints. This is handled by passing to the continuous relaxation of the discrete problem. That is, we no longer restrict the n_i 's to be natural numbers, but allow them to take on real values as well. It suffices to maximize

$\log W(\mathbf{n})$ subject to the constraints since the map $x \to \log\left(x\binom{M}{N}\right)$ is monotonic and thus $\pi(\mathbf{n})$ and $\log W(\mathbf{n})$ will be maximized for the same values of $\mathbf{n}$.

The problem becomes to maximize

$$\log W(\mathbf{n}) = \sum_{i=1}^{m} \log \binom{g_i}{n_i} = \sum_{i=1}^{m} \log \frac{g_i!}{\Gamma(n_i+1)\Gamma(g_i-n_i+1)} =$$

$$\sum_{i=1}^{m} \log g_i! - \log\Gamma(n_i+1) - \log\Gamma(g_i-n_i+1)$$

subject to the constraints

$$\sum_{i=1}^{m} n_i = N, \qquad \sum_{i=1}^{m} n_i \epsilon_i = E, \quad 0 \le n_i \le g_i$$

where Γ is the Gamma function, a generalization of the factorial to real and complex values which is being used because n_i can take on any real value, and thus the usual formula $\binom{g_i}{n_i} = \frac{g_i!}{n_i!(g_i-n_i)!}$ is not valid. For natural numbers n, $n! = \Gamma(n+1)$.

The quantity $\log W(\mathbf{n})$ is proprotional to a quantity known as the Boltzmann entropy and the Fermi-Dirac distribution is the distribution that maximizes the Boltzmann entropy among all of those satisfying the constraints.

Since $\sum_{i=1}^{m} \log g_i!$ does not depend on $\mathbf{n}$, it suffices to maximize

$$-\sum_{i=1}^{m} \log\Gamma(n_i+1) + \log\Gamma(g_i-n_i+1)$$

subject to the constraints.

The expression appears unwieldy but can be simplified with Stirling's approximation in the form

$$\log\Gamma(x+1) = x\log x - x + O(\log x)$$

by replacing $\log\Gamma$ terms with Stirling's approximation, ignoring error terms, simplifying, and then ignoring terms that do not depend on $\mathbf{n}$ we end up with the problem of maximizing

$$-\sum_{i=1}^{m} n_i \log(n_i) + (g_i - n_i)\log(g_i - n_i)$$

again subject to the constraints

$$\sum_{i=1}^{m} n_i = N, \qquad \sum_{i=1}^{m} n_i \epsilon_i = E, \quad 0 \leq n_i \leq g_i$$

(As an aside, note that we haven't justified ignoring the error terms. It's common in physics to treat error terms as negligible when deriving a mathematical model. There are often implicit assumptions baked into the treatment of certain terms as negligible. Agreement with experiment remains the ultimate test of physical theory. It will serve a mathematician well to become comfortable working non-rigorously like a phycisist; one can cover more ground that way. A vital skill for a mathematician is then to be able to convert a loose argument into a rigorous one or otherwise to identify when obstructions make this difficult or impossible. See here for an excellent post on this matter by an immeasurably more skilled and knowledgeable mathematician.)

This constrained optimization problem can be solved by the method of Lagrange multipliers. (Technically, since some of the constraints are inequalities, one actually applies the Karush Kuhn Tucker conditions.)

one can then find the optimal value for $\mathbf{n}$ to be

$$\hat{n}_i = \frac{g_i}{1 + e^{\alpha + \beta \epsilon_i}} \quad i = 1 \dots m$$

for real constants α and β for which there is no closed form solution except in special cases. These constants have a physical interpretation that we will not discuss here. The probability $\pi(\epsilon | N, E)$ that a state s is occupied given its energy level and for fixed N and E is then taken to be

$$\pi(\epsilon | N, E) = \frac{\hat{n}_i}{g_i} = \frac{1}{1 + e^{\alpha + \beta \epsilon_i}}$$

This expression may look familiar if you've seen logistic regression before. Observe that it has no dependence on the values of the individual g_is. The resulting probability distribution is known as the Fermi-Dirac distribution.

You may not be entirely satisfied with the above derivation. We haven't really explained why we should expect this approach to yield the correct distribution. You may may be more satisfied by the following interpretation.

Through a series of simplifications we see that the quantity we are maximizing satisifes the equivalence

$$-\sum_{i=1}^{m} n_i \log n_i + (g_i - n_i) \log (g_i - n_i) =$$

$$-g_i \sum_{i=1}^{m} \frac{n_i}{g_i} \log \frac{n_i}{g_i} + (1 - \frac{n_i}{g_i}) \log \left(1 - \frac{n_i}{g_i}\right) + \log g_i$$

maximizing it under the constraints is then equivalent to maximizing

$$-\sum_{i=1}^{m} \frac{n_i}{g_i} \log \frac{n_i}{g_i} + (1 - \frac{n_i}{g_i}) \log \left(1 - \frac{n_i}{g_i}\right)$$

letting $p_i = \frac{n_i}{g_i}$ and interpreting p_i to be the probability that a state s of energy level ϵ_i is occupied for fixed N and E, the above expression becomes

$$-\sum_{i=1}^{m} p_i \log p_i + (1 - p_i) \log (1 - p_i)$$

you may recognize this expression as the Shannon entropy of a discrete probability distribution.

(If you've not yet read Claude Shannon's landmark paper *A Mathematical Theory of Communication* [5], please read it. If you've read it before, perhaps read it again (I plan to shortly). You can find it online here. It is accessible to anyone capable of following this post and filled with important insights and deep wisdom.)

We thus see that the Fermi-Dirac distribution is the distribution that maximizes the Shannon entropy under the constraints that the total number of particles N and the total energy E are fixed. By the principle of maximum entropy, we see that the Fermi-Dirac distribution makes the fewest assumptions about the true distribution given the constraints that N and E are fixed at particular values. The Fermi-Dirac distribution thus has a theoretical elegance to it and its agreement with experiment may give one the (justified in my opinion) impression that statistical physics is rooted in profound truths about the world.

This direct connection between Shannon entropy and Boltzmann entropy was first made by E.T. Jaynes [6] who developed an interpretation of statistical physics as an application of Bayesian inference and information theory. His posthumously published book Probability theory: The logic of science [7] offers a fascinating look into his thought on the foundations of Bayesian probability and inference.

Review of binary classification

Now let's give a quick review of the binary classification problem. I think the best references for this sort of material are the book The Elements of Statistical Learning by Hastie, Tibshirani, and Friedman [8]; and for those who use Python, the excellent Sci-kit learn documentation [9].

Suppose we are given a set $\mathbf{X}$ consisting of points in a p dimensional Euclidean space $\mathbb{R}^p$. An example of such points could be vectors consisting of pixel intensities from 4000×4000 grayscale images. In this case $p = 4000 \times 4000 = 16,000,000$. (The individual coordinates of the vectors are

often called features. We might say elements of the image dataset above each have 16 million features.) There is an unknown function $f : \mathbf{R}^p \to \{0, 1\}$ which maps $\mathbf{x}$ to an associated target y that is either 0 or 1 depending on whether $\mathbf{x}$ belongs to some class of interest. An example could be a function mapping 4000×4000 pixel grayscale images to 1 if an image contains a cat and 0 if it does not. We then consider the following problem:

Given a subset $\mathcal{X}$ consisting of n points from $\mathbf{X}$ along with a set of n corresponding labels $\mathbf{y}$ for the points in $\mathcal{X}$, find a function $\hat{f} : \mathbb{R}^p \to \mathbb{R}$ from a family of functions $\mathcal{F}(\boldsymbol{\alpha})$ parametrized by $\boldsymbol{\alpha} \in \mathbb{R}^k$ and a decision rule $r : \mathbb{R} \to \{0, 1\}$ such that the composition $r \circ \hat{f} : \mathbb{R}^p \to \{0, 1\}$ closely approximates the unknown function f. For example, if given a collection of 2 million 4000×4000 pixel grayscale images, each labeled as to whether the image contains a cat, we are essentially trying to find a function that is able to identify whether any 4000×4000 pixel grayscale image contains a cat. The set $\mathcal{X}$ is referred to as the training data.

(Note that we are only discussing a special case of the classification problem where there is a fixed but unknown function f. Often it's actually the case that there is uncertainty in the output of f, so that the same point $\mathbf{x}$ may have different labels in different cases, perhaps dependent on some unobserved variables or due to inherent randomness. For the cat pictures example, it's somewhat sensible to say that f is fixed. If one were to try to predict movement in the stock market from past movements, f is anything but fixed. Restricting to the case of fixed f simplifies the discussion in some places while doing no real harm to the generality of the results.)

A family of functions $\mathcal{F}(\boldsymbol{\alpha})$, along with an algorithm which when given a set of training data attempts to find a good choice of $\boldsymbol{\alpha}$ to determine the function $\hat{f}$, is called a classification model. We say a classification model is fitting a function $\hat{f}$ based on the training data $\mathcal{X}$ and this process is called training. After fitting we say we have trained a classification model. Examples of classification models include logistic regression, decision trees, support vector machines, random forests, gradient boosted tree ensembles, and neural networks.

For some classification models such as logistic regression and random forests, the function $\hat{f}$ maps points in $\mathbf{X}$ to properly calibrated probabilities that the associated objects belong to the category of interest. Such methods can produce functions which when given an image return a probability that it contains a cat. If you were to look at many such images where the predicted probability is 0.65, you should expect that approximately 65% of them actually do contain a cat. The decision function r is then determined by choosing a probability threshold p such that $r(x) = 1$ if $x \geq p$ and equals 0 otherwise.

Other methods such as support vector machines and gradient boosted trees do not produce functions $\hat{f}$ that return calibrated probabilities. Gradient boosted trees models produce predicted values in the interval $[0, 1]$ but they tend to cluster at the tails, making predictions that either under or over estimate the true probabilities. A support vector machine identifies a hyperplane in $\mathbb{R}^p$ which separates points with label 1 from points with label 0. The function $\hat{f}$ gives the signed distance

between a point and the hyperplane so the decision rule r identifies the side of the hyperplane on which a point lies.

In either case, it is customary to arrange things such that if $\hat{f}(x_1) > \hat{f}(x_2)$, then the classifier believes x_1 is more likely to be in the class of interest (e.g. to be a picture of a cat) than x_2 and the decision rule r is based on a cutoff value x_0 such that $r(x) = 1$ if $x \geq x_0$ and equals 0 otherwise. From now on we will assume this to be the case.

A natural problem is then: given a classification model which does not produce calibrated probability scores and a function $\hat{f}$ fit by this model, to identify an additional function $c : \mathbb{R} \to [0, 1]$ such that $c \circ \hat{f} : \mathbb{R}^p \to [0, 1]$ produces correctly calibrated probabilities. Typically c is fit with another classification model which takes the scores returned by $\hat{f}$ on a set of labeled points as its training data. This problem is known as the problem of probability calibration and it is the primary problem considered by the authors of [1].

Probability calibration through Fermi-Dirac statistics

The authors of [1] consider a classification model $\mathcal{M}$ which has fit a function $\hat{f}$ which does not return calibrated probabilities. Their formulation of the problem corresponds to a Fermi gas with no degeneracies, that is, one where there is only one state per energy level. The above derivation of the Fermi-DIrac distribution is not applicable in this case, so I'm going to use an alternative formulation to work around this. For those interested, see the original paper and particularly its appendix for information on how they formulate the problem.

Given a set of labeled training data $\mathcal{X}$ with n elements and that is disjoint from the training data that was used to fit $\hat{f}$, to each point $\mathbf{x} \in \mathcal{X}$ they associate the score $\hat{f}(\mathbf{x})$. The scores are then sorted in order from highest to lowest. Scores corresponding to points for which the classifier is more confident belong to the class of interest appear before points for which it is less confident.

Scores are then replaced with their ranks within this sorted list, so that each point in $\mathcal{X}$ is mapped to a natural number between 1 and n such that if $\mathbf{x}_i$ is mapped to a lower number than $\mathbf{x}_j$, then the classifier is more confident that $\mathbf{x}_i$ belongs to the class of interest. The authors assume that ties are broken uniformly at random.

Ranks for unseen datapoints $\mathbf{x}$ not in $\mathcal{X}$ can be computed by calculating $x = \hat{f}(\mathbf{x})$, situating it among the sorted scores from

$$\text{Scores} = \left\{ \hat{f}(\mathbf{x}_i) : \mathbf{x}_i \in \mathcal{X}_1 \right\}$$

and assigning it the rank of the nearest score in the set Scores, perhaps making a determination to always assign the smaller rank if $\hat{f}(\mathbf{x})$ is at the midpoint of two neighboring scores.

In this way they define a function

$\mathrm{rank}: \mathbf{X} \to [1, \ldots, n]$.

One can construct a formal mapping onto the problem investigated through Fermi-Dirac statistics in the following way. Assume $\mathcal{X}$ was a random sample from the set $\mathbf{X}$ of otherwise unlabeled data. Each datapoint $\mathbf{x}$ in $\mathbf{X}$ is considered to be a possible state for a fermion. If $\mathbf{x} \in \mathbf{X}$ then let the energy level associated to the state $\mathbf{x}$ be defined as $\mathrm{rank}(\mathbf{x})$.

The state $\mathbf{x}$ is considered to be occupied by a fermion if the true label corresponding to the point $\mathbf{x}$ is 1 (e.g. grayscale images containing cats are occupied by fermions), otherwise the state is considered to be unoccupied.

Each state $\mathbf{x}$ can be occupied by at most one fermion since each datapoint is assumed to have a unique label of either 0 or 1 and thus the Pauli exclusion principle holds. We assume that $\mathbf{X}$ is sufficiently large compared to $\mathcal{X}$ that each possible rank will appear enough times to ensure the degeneracies are large enough to apply the above derivation of the Fermi-Dirac distribution.

The total number of fermions N in the Fermi gas is then

$$N = \sum_{\mathbf{x} \in \mathbf{X}} [f(\mathbf{x}) = 1]$$

and the total energy E is

$$E = \sum_{\mathbf{x} \in \mathbf{X}} [f(\mathbf{x}) = 1] \, \mathrm{rank}(\mathbf{x})$$

where $[\star]$ is the Iverson bracket notation defined by $[\star] = 1$ if $\star$ is true, otherwise $[\star] = 0$. Recall that f is the unknown function that maps each $\mathbf{x} \in \mathbf{X}$ to its true label y.

If N and E are known, then the probability $\pi(k|N, E)$ that a datapoint s has true label 1 given that $\mathrm{rank}(s) = k$ can be estimated by the Fermi-Dirac distribution. The authors stress that this is not necessarily the true probability that the true label is 1 given the rank, but is the distribution that makes the minimal number of assumptions given the constraints.

Although N and E are not known, they can be estimated based on the labeled dataset $\mathcal{X}$ for which the total number of fermions and the total energy *can* be calculated. One simply extrapolates the average number of fermions per datapoint and the average energy per datapoint within $\mathcal{X}$ to all of $\mathbf{X}$.

Calibrated probability scores can then be computed through the formula

$$\text{Prob}\left[f(s)=1\right] \approx \frac{1}{1+e^{\alpha+\beta\,\text{rank}(s)}}$$

Those who are familiar with logistic regression may feel something is a little suspicious. Haven't we just fit a logistic regression model using rank transformed features? Let's see what the authors have to say about it

Some methods, however, explicitly model the posterior probability of their classification, for example using logistic regression or Platt scaling methods (43), performing a logistic transformation of chosen features in the former or of a classifier score in the latter, into an output probability. While such transformations make intuitive sense and work well for some applications, they are heuristic methodologies. Our approach is different from the abovementioned methods on two counts: On the one hand, our logistic transformation transforms the ranks (not features or scores) assigned by a classifier to items in a test set into a probability; on the other hand, the logistic transformation is not postulated as an ad hoc transformation but results from the maximum-entropy principle and as such is the least-biased distribution given the information at hand. In other words, ours is the most parsimonious calibrated class distribution, and, in the absence of additional information, should be preferred to other methods.

OK, they've fit a function

$$p(r) = \frac{1}{1+e^{\alpha+\beta r}}$$

mapping ranks to probabilities that has the same functional form as what would be produced by a logistic regression model. Are the parameters α and β somehow different from the ones logistic regression would fit? Does this mean that logistic regression somehow produces parameters α and β that are suboptimal because it is only a heuristic methodology while their method produces better parameters because it is the least biased distribution given the information at hand? Have statisticians really been doing it wrong for all these years? I think it's really unfortunate that this fact is not more widely known but it turns out that logistic regression is actually characterized by producing the distribution that is the least-biased given the information at hand.

The Max Entropy Classifier is Logistic Regression

Wait, one may say: "I don't recall reading anything about logistic regression producing a maximum entropy distribution in *The Elements of Statistical Learning* nor in Agresti's classic textbook *Categorical Data Analysis* [10] which contains hundreds of pages devoted to logistic regression and its applications."

I would be astonished if the authors of these books were not aware of this fact about logistic regression. It's not entirely clear why no mention is made of it. The mathematical details are certainly not beyond the level of things that are presented in the first of these books. Each book is already quite long, but I think it could at least have merited an exercise.

In any case, one can find a lucid exposition in a 1996 paper by Berger et al [11] entitled *A Maximum Entropy Approach to Natural Language Processing*. There is a nice 2011 article by John Mount, which can be found on github here [12] and from which I first learned the details of this equivalence.

Suppose we have a labled dataset $\mathcal{X}$ with elements from a larger set of mostly unlabled data $\mathbf{X} \subset \mathbb{R}^p$. We seek a function $\pi : \mathbf{X} \to [0,1]$ such that $\pi(\mathbf{x})$ gives an estimate that the true label corresponding to $\mathbf{x} \in \mathbf{X}$ is 1 for any $\mathbf{x} \in \mathbf{X}$.

Following either of the above references one show that the max entropy classifier is equivalent to logistic regression. The max entropy classifier can be derived by considering π satisfying the following constraints

$$0 \leq \pi(x) \leq 1, \ \ \forall \mathbf{x} \in \mathbf{X}$$

$$\sum_{\mathbf{x} \in \mathcal{X}} \pi(\mathbf{x}) x_i = \sum_{\mathbf{x} \in \mathcal{X}} [f(\mathbf{x}) = 1] x_i, \ \ i = 1 \dots p$$

where x_i denotes the ith coordinate of $\mathbf{x}$.

(The first constraint ensures that π produces valid probabilities. The second asks that the sum in each particular coordinate over all values in the training data with label 1 is equal to the expected sum over that coordinate under the distribution π).

From all functions π satisfying the constraints, one chooses that which maximizes the Shannon entropy

$$H(\pi, \mathcal{X}) = -\sum_{\mathbf{x} \in \mathcal{X}} \pi(\mathbf{x}) \log(\pi(\mathbf{x})) + (1 - \pi(\mathbf{x})) \log(1 - \pi(\mathbf{x}))$$

Note that when $p = 1$ these are equivalent conditions to those we used to derive the Fermi-Dirac distribution. The fitted function is thus exactly the same.

Now consider the standard method of deriving logistic regression, let's work in the case where $p = 1$ for simplicity. One assumes that $\pi(x)$ follows the functional form

$$\pi(x) = \frac{1}{1 + e^{\alpha + \beta x}}$$

and from there one attempts to maximize the log-likelihood

$$\mathcal{L}(\mathcal{X}|\alpha,\beta) = \sum_{x\in\mathcal{X}} [f(x)=1]\log\pi(x) + [f(x)=0]\log(1-\pi(x)) =$$

$$-\sum_{x\in\mathcal{X}} [f(x)=1]\log(1+e^{\alpha+\beta x}) + [f(x)=0]\log(1+e^{-\alpha-\beta x})$$

(The signs of α and β have been flipped from the standard presentation of logistic regression in order to make things more closely match the presentation of Fermi-Dirac statistics given above.)

This is an unconstrained optimization problem, we thus proceed by finding values α and β for which the gradient of $\mathcal{L}$ equals 0.

Through some simplifications we find

$$\frac{\partial\mathcal{L}}{\partial\alpha} = -\sum_{x\in\mathcal{X}} (1-\pi(x))\,[f(x)=1] + \pi(x)[f(x)=0]$$

$$\frac{\partial\mathcal{L}}{\partial\beta} = -\sum_{x\in\mathcal{X}} x(1-\pi(x))\,[f(x)=1] + x\pi(x)[f(x)=0]$$

from which it's possible to derive the constraint

$$\sum_{x\in\mathcal{X}} \pi(x)x = \sum_{x\in\mathcal{X}} [f(x)=1]x$$

By working through the equations for each method, it's possible to show that standard method produces the same α and β as the max entropy method. We leave the details out for brevity, see the aforementioned article by Berger et al for a more thorough proof.

OK, so it turns out that the authors do seem to be fitting a logistic regression of some kind. But what of their claim that their method differs from logistic regression because "our logistic transformation transforms the ranks (not features or scores) assigned by a classifier to items in a test set into a probability"?

I admit that in all of the occasions I've reached for Platt scaling, I never considered whether it might be beneficial to rank transform the predicted scores before fitting logistic regression. This doesn't mean they are not fitting a logistic regression though, it only means they are applying a feature transform. A crude version could be implemented in Python with numpy and Scikit-learn like this

```
import numpy as np
from sklearn.base import BaseEstimator
from sklearn.base import TransformerMixin
from sklearn.linear_model import LogisticRegression
from sklearn.pipeline import Pipeline
```

```
class RankTransformer(BaseEstimator, TransformerMixin):
    """Apply rank transform to features

    Warning: not production ready
    """
    def fit(self, X, y=None):
        self.train_sorted = np.sort(X, axis=0)
        return self

    def transform(self, X, y=None):
        n, p = X.shape
        assert p == self.train_sorted.shape[1]
        ranks = np.hstack(
            [
                np.searchsorted(self.train_sorted[:, i], X[:, i]).reshape(n, 1)
                for i in range(p)
            ]
        )
        return ranks + 1
```

Implementing a Fermi-Dirac classifier is then as simple as

```
fermi_dirac_classifier = Pipeline(
    [
        ('rank_transform', RankTransformer()),
        ('logit', LogisticRegression(penalty=None, solver="newton-cg"))
    ]
)
```

The authors go on to discuss a method they call FiDEL which amounts to the method of pooling probabilities predicted by different methods by taking the geometric mean of the resulting odds ratios. A discussion of the merits of this method in comparison to others can be found in a highly relevant paper by Satopää et al[13]. In the case of FiDEL, the scores output by different classifiers are converted to calibrated probabilities through the author's physics inspired variant of Platt scaling, and then combined in this way.

The most interesting parts of [1] connect the physical interpretations of the parameters α and β within a Fermi gas with the interpretation of these parameters in binary classification.

Conclusion

OK, so how does this happen? Why were a group of physicists who are obviously skilled and mathematically competent unable to detect that they'd rediscovered known methods? How is it that none of the reviewers for their PNAS paper were able to point this out to them? I think part of the answer lies in how vast the sphere of human knowledge really is. It is beyond anyone to keep up with the details of more than small portions of it. Research communities focus on their particular sets of problems and communication between different research communities is often weak to non-existent. This is not out of laziness, incompetence, or malevolence, but because the difficulty of such communication tests our limits. Just keeping on top of ones own field and publishing consistently is challenging enough. How to properly review cross-disciplinary research is a difficult problem.

Transitioning from one research field to another is difficult. Over ten years ago my goal of having a pure math career was upended due to health issues. As I recovered, my interests switched to machine learning simply because it was an in demand field with career potential outside of the brutal competition for faculty jobs. I had some bad takes early on. I was skeptical of the potential of deep learning because it seemed ad-hoc and unprincipled. In just a short time I was proven very wrong. It took over six years from when I changed focus before I felt capable of developing novel and useful algorithms.

Had I been asked to review this paper, I would not have recommended rejection. Certainly I would have asked for major revisions, but ultimately I think this is worthwhile work. I would have pointed out that they were in fact fitting a logistic regression model and basically made them aware of all of the contents of this post. I would have asked them to rewrite the paper with this awareness in mind; to rewrite it as a discussion of the connections between statistical physics and logistic regression offering a means to bring physical intuition to binary classification problems. The paper could have been a bridge between the statistical physics and statistical machine learning communities, allowing practioners of each field insight into the other through discussion of a shared problem. Also, I had never considered the possibility of rank transforming the features before Platt scaling and the authors make some compelling arguments for it. Like most techniques it will probably be beneficial in some cases and harmful in others with empirical performance being the final judge, but it's an interesting tool to keep in mind.

Judgement: *The paper is thought provoking in its own right regardless of whether the methods are novel and is worth a read.*

References

[1] Kim SC, Arun AS, Ahsen ME, Vogel R, Stolovitzky G. The Fermi-Dirac distribution provides a calibrated probabilistic output for binary classifiers. Proc Natl Acad Sci U S A. 2021 Aug

24;118(34):e2100761118. doi: 10.1073/pnas.2100761118. PMID: 34413191; PMCID: PMC8403970.

[2] Gibbs, Josiah Willard (1902). Elementary Principles in Statistical Mechanics. New York: Charles Scribner's Sons.

[3] Dirac, Paul A. M. (1926). "On the Theory of Quantum Mechanics". Proceedings of the Royal Society A. 112 (762): 661–77.

[4] Fermi, Enrico (1926). "Sulla quantizzazione del gas perfetto monoatomico". Rendiconti Lincei (in Italian). 3: 145–9., translated as Zannoni, Alberto (1999-12-14). "On the Quantization of the Monoatomic Ideal Gas". arXiv:cond-mat/9912229

[5] Shannon, C.E. (1948), A Mathematical Theory of Communication. Bell System Technical Journal, 27: 379-423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

[6] E. T. Jaynes, Information theory and statistical mechanics. Phys. Rev. 106, 620–630 (1957).

[7] Jaynes, E. T., & Bretthorst, G. L. (2003). Probability theory: The logic of science. Cambridge, UK: Cambridge University Press.

[8] Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). The elements of statistical learning: data mining, inference, and prediction. 2nd ed. New York: Springer.

[9] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

[10] Agresti, A. (2013) Categorical Data Analysis. 3rd Edition, John Wiley & Sons Inc., Hoboken.

[11] Adam L. Berger, Vincent J. Della Pietra, and Stephen A. Della Pietra. 1996. A maximum entropy approach to natural language processing. Comput. Linguist. 22, 1 (March 1996), 39–71.

[12] Mount, J. (2011). The equivalence of logistic regression and maximum entropy models. https://github.com/WinVector/Examples/blob/main/dfiles/LogisticRegressionMaxEnt.pdf

[13] Satopää, Ville & Baron, Jonathan & Foster, Dean & Mellers, Barbara & Tetlock, Philip & Ungar, Lyle. (2014). Combining multiple probability predictions using a simple logit model. International Journal of Forecasting. 30. 344–356. 10.1016/j.ijforecast.2013.09.009.

Albert Steppi
first dot last at gmail

Musings on math, technology, and the fate of industrial civilization.