Logistic Regression

Remember conditional probability?

Definition 1 (Kolmogorov)**.**

$$\mathbf{Pr}(A \mid B) = \frac{\mathbf{Pr}(A \cap B)}{\mathbf{Pr}(B)}$$

Definition 2 (axiom)**.**

$$\mathbf{Pr}(A \cap B) = \mathbf{Pr}(A \mid B)\,\mathbf{Pr}(B)$$

Example 1 (mostly reliable)**.** *Suppose a specific test protocol has a 20% false negative rate and a 1% false positive rate. If you test positive, what is the probability you are positive?*

Suppose first that the incident rate is 50%. That it is 1%.

Solution. Show with measuring squares, then with (Kolmogorov) definition. *solution/*

Example 2 (less reliable)**.** *Suppose a specific test protocol has a 10% false negative rate and a 50% false positive rate. If you test positive, what is the probability you are positive?*

Suppose first that the incident rate is 50%. That it is 1%.

Solution. Show with measuring squares, then with (Kolmogorov) definition. *solution/*

Example 3 (generating points)**.** *Given a point and a gaussian distribution, what is the probability that a point is produced at a given location? Within a given interal?*

Solution. Show with picture. *solution/*

Example 4 (validating points)**.** *Given a set of points and a new point, what is the probability that the new point is part of that distribution?*

Solution. This is actually logistic regression.

- Is this obvious?
- Is logistic regression a linear model? Why or why not?

Theorem 1. *Bayes*

$$\mathbf{Pr}(H \mid D) = \frac{\mathbf{Pr}(D \mid H)\,\mathbf{Pr}(H)}{\mathbf{Pr}(D)}$$

Proof. From axiom definition and symmetry. □

Linear model: $y = \beta x + \beta_0$.

Logistic regression = learn $\mathbf{Pr}(H \mid D)$ or at least $\mathbf{Pr}(H \mid x \in D)$.

Usually one now simply shows

$$\mathbf{Pr}(y \mid x) = \frac{1}{1 + e^{-(\beta x + \beta_0)}}$$

which is the inverse logit function, β, β_0 are to be learned.

Recall: logit is $\sigma(x) = 1/(1 + e^{-x})$ and is defined by

$$logit(p) = \sigma^{-1}(p) = \ln\left(\frac{p}{1-p}\right)$$

Example 5. *Making bread during lockdown.*
H is "made a good loaf of bread".
D is vector of technique details (what I did).
Want $\mathbf{Pr}(H \mid D)$, the posterior probability (probabilité à posteriori)

This is our signal to look at Bayes theorem.

$$\mathbf{Pr}(H \mid D) = \frac{\mathbf{Pr}(D \mid H)\,\mathbf{Pr}(H)}{\mathbf{Pr}(D)}$$

Recall that $\mathbf{Pr}(D \mid H)$ is the likelihood (vrasemblance), $\mathbf{Pr}(H)$ and $\mathbf{Pr}(D)$ are the prior (or marginal) probabilities (probabilités à priori).

We also call $\mathbf{Pr}(D)$ the evidence.

$\mathbf{Pr}(H)$ we observe. But the others? Especially $\mathbf{Pr}(D)$ we have no clue.

So we introduce $\mathbf{Pr}(\overline{H})$.

$$\mathbf{Pr}(\overline{H} \mid D) = \frac{\mathbf{Pr}(D \mid \overline{H})\,\mathbf{Pr}(\overline{H})}{\mathbf{Pr}(D)}$$

Divide, the $\mathbf{Pr}(D)$'s cancel. We'll call these odds (cote in French).

$$C(H \mid D) = \frac{\mathbf{Pr}(H \mid D)}{\mathbf{Pr}(\overline{H} \mid D)} = \frac{\mathbf{Pr}(D \mid H)\,\mathbf{Pr}(H)}{\mathbf{Pr}(D \mid \overline{H})\,\mathbf{Pr}(\overline{H})} = \frac{\mathbf{Pr}(D \mid H)}{\mathbf{Pr}(D \mid \overline{H})}\,C(H)$$

A linear model would look like $y = \beta x + \beta_0$. We don't have that yet. But we could take logs.

$$\ln(C(H \mid D)) = \ln\left(\frac{\mathbf{Pr}(D \mid H)}{\mathbf{Pr}(D \mid \overline{H})}\right) + \ln(C(H))$$

The first term we call the log likelihood.
The second term is basically constant, the log prior.

Assume the log likelihood is basically a linear function of D. Then we get

$$\ln(C(H \mid D)) = \beta D + \beta_0$$

And we can learn β, β_0, then take exponents to get to the quantity $\mathbf{Pr}(H \mid D)$ that we wanted at the start. *solution/*

There's more work to do. We need to do some work to understand in detail how to pass this to an optimiser, what the right loss function is for gradient descent. We're not going to do that here.

Lemma 1 (Aside, only discuss in class if questioned.)**.** *We can recover* $\mathbf{Pr}(H)$ *from* $C(H)$.

Proof.

$$\begin{aligned} C(H) &= \frac{\mathbf{Pr}(H)}{\mathbf{Pr}(\overline{H})} = \frac{\mathbf{Pr}(H)}{1 - \mathbf{Pr}(H)} \\ \mathbf{Pr}(H) &= (1 - \mathbf{Pr}(H))C(H) \\ &= \frac{1 + C(H)}{1 + C(H)}(1 - \mathbf{Pr}(H))C(H) \\ &= \frac{(1 + C(H))(1 - \mathbf{Pr}(H))C(H)}{1 + C(H)} \end{aligned}$$

$$\begin{aligned} (1 + C(H))(1 - \mathbf{Pr}(H)) &= \left(1 + \frac{\mathbf{Pr}(H)}{\mathbf{Pr}(\overline{H})}\right)(1 - \mathbf{Pr}(H)) \\ &= \frac{\mathbf{Pr}(\overline{H}) + \mathbf{Pr}(H) - \mathbf{Pr}(H) - \mathbf{Pr}(H)^2}{\mathbf{Pr}(\overline{H})} \\ &= \frac{\mathbf{Pr}(\overline{H}) + \mathbf{Pr}(H) - \mathbf{Pr}(H)(\mathbf{Pr}(\overline{H}) + \mathbf{Pr}(H))}{\mathbf{Pr}(\overline{H})} \\ &= \frac{1 - \mathbf{Pr}(H)}{\mathbf{Pr}(\overline{H})} \\ &= \frac{1 - \mathbf{Pr}(H)}{1 - \mathbf{Pr}(H)} \\ &= 1 \end{aligned}$$

□

Review of this review:

- Logistic regression is a linear model.
- It's based on Bayes Theorem. It gives us probabilities based on observed data.
- It's simple to interpret.

The point of this course is going to be to understand two other separation algorithms: SVM and ANN.

Both have the property that they can be linear or transform their problem to find non-linear separators.

One of the goals of this course is that you **don't** think of your job as trying things until one works.

SVM

Quick introduction to the motivation for SVM: maximum margins.

Reminder for next time: projects, git repos. Test your python installations (or the ones the school provides).