20. Linear Predictors (part 3: logistic regression)

Monday, February 19, 2024 5:25 PM

It's not really regression, it's classification where instead of outputting a label 0 or 1 (or ±1), we output a probability reflecting our confidence in a label

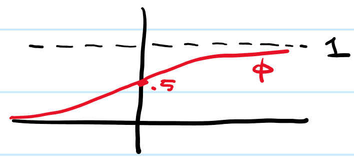
not necessarily a "true prob.", but at least in $[0,1]$

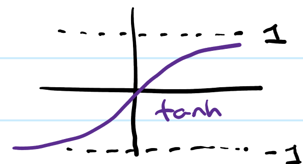
$\mathcal{H} = \phi \circ L_d$ with

"sigmoid"

$$\phi(z) := \frac{1}{1+e^{-z}} = \frac{e^z}{1+e^z} = \frac{1}{2}(\tanh(z)+1)$$

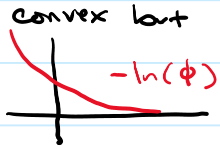
often denoted σ





$\tanh(z) = \frac{e^{2z}-1}{e^{2z}+1}$

- ϕ isn't convex but $-\ln(\phi(z))$ is (proof:)



$$f(z) = -\ln(\phi(z)) = \ln\left(\frac{1}{\phi(z)}\right) = \ln(1+e^{-z})$$

$$\left(\frac{f}{g}\right)' = \frac{f'g - g'f}{g^2}$$

so $f'(z) = \frac{-e^{-z}}{1+e^{-z}}$ and $f''(z) = \frac{e^{-z}(1+\cancel{e^{-z}}) + e^{-z}(-\cancel{e^{-z}})}{(1+e^{-z})^2} = \frac{e^{-z}}{(1+e^{-z})^2} > 0$ so convex.

- $\phi^{-1}(p) = \ln\left(\frac{p}{1-p}\right) =:$ logit(p) if $p \in (0,1)$
- $2\phi(z) - 1 \approx \text{sign}(z)$ if $|z| \gg 1$
- $\phi(z) = 1 - \phi(-z)$ symmetry

Motivation / justification for loss function

Assume $D_{y|x}$ is given by $y = \begin{cases} +1 & \text{w.p. } \sigma(\langle w^*, x\rangle) \\ -1 & \text{w.p. } 1-\sigma(\langle w^*, x\rangle) \end{cases}$

or concisely, $\mathbb{P}[Y=y \mid w^*, x] = \sigma(y \cdot \langle w^*, x\rangle)$ since $\phi(z) = 1-\phi(-z)$

an example of a Generalized Linear Model (GLM)

$$\mathbb{P}[Y=y \mid x] = g^{-1}(\langle w^*, x\rangle)$$

["General" Linear Model just means several columns and Frobenius norm]

then if we have m iid observations,

the likelihood is $\mathbb{P}[(y_i)_{i=1}^m \mid w, (x_i)_{i=1}^m] = \prod_{i=1}^m \mathbb{P}[y_i \mid w, x_i]$

and the maximum likelihood estimator (MLE) maximizes this likelihood (or minimize negative log likelihood)

$$w_{MLE} \in \underset{w}{\text{argmin}} -\sum_{i=1}^m \ln(\sigma(y_i \langle w, x_i\rangle))$$

$$= \underset{w}{\text{argmin}} + \sum_{i=1}^m \ln(1 + \exp(-y_i \langle w, x_i\rangle))$$

20a. Linear Predictors (part 3: logistic regression)

Monday, February 26, 2024 9:38 AM

So if we choose our loss function $\ell(\hat{y}, y) = \log(1 + e^{-y \cdot \hat{y}})$

then ERM is just MLE

(and logistic regression is well-understood; see ch. 24 for more on MLE)

Random numerical issues

Compute $f(a) = \ln(1 + e^a)$ for $a = 400 \cdot \ln(10)$ (so $e^a = 10^{400}$)

it should be $f(a) \approx a$ since $1 + e^a \approx e^a$ and $\ln(e^a) = a$

but numerically, the intermediate term e^a produces overflow

Instead:

$$\ln(1 + e^a) = \ln(e^a(e^{-a} + 1)) = a + \ln(\underbrace{1 + e^{-a}}_{\text{underflow}})$$

and calculate

$$\ln(1 + e^{-a}) = \text{log1p}(e^{-a})$$

which uses Taylor series

$$\log(1 + x) = x - x^2/2 + x^3/3 - \ldots$$

Similarly: numpy.logaddexp

for $\log(e^a + e^b)$