TAM 598 Lecture 16:

Classification

Announcements:

- HW 4 covers lectures 13-16; due on Fri Apr 4
  ↑ updated!

I. Classification / logistic regression

observations $\underset{\sim}{x}_{1:n} = (\underset{\sim}{x}_1, \underset{\sim}{x}_2, \ldots, \underset{\sim}{x}_n)$

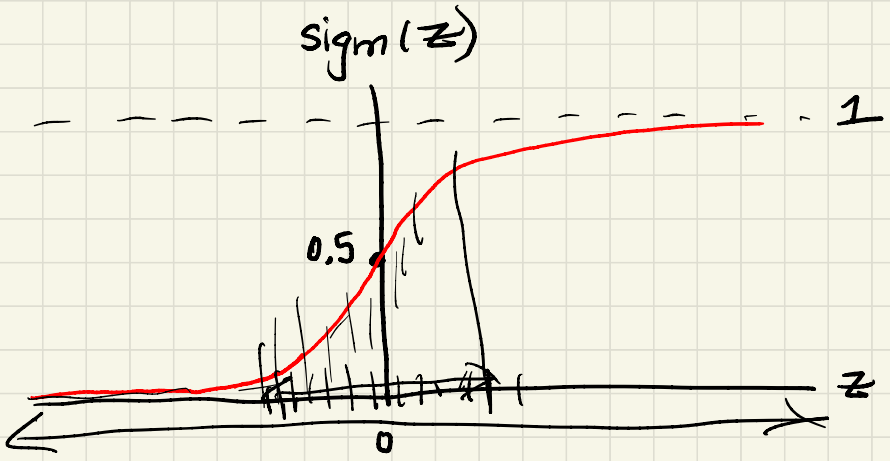
targets $y_{1:n} = (y_1, \ldots, y_n)$ ← discrete labels

consider binary classification, where $y = 0$ or $y = 1$.

probability that $y = 1$ conditioned on x:

$$p(y=1 \mid \underset{\sim}{x}, \underset{\sim}{w}) = \text{sigm}\left(\sum_{j=1}^{m} w_j \phi_j(\underset{\sim}{x})\right) = \text{sigm}\left(\underset{\sim}{w}^T \phi(\underset{\sim}{x})\right)$$

where: w_j's are weights, $\phi_j(\underset{\sim}{x})$ are basis functions, and "sigm" is the sigmoid function $\text{sigm}(z) = \dfrac{1}{1+e^{-z}}$



logistic regression → a generalized linear model pushed through a sigmoid function so it is mapped to $[0,1]$

then $p(y=1) = \text{sigm}\left(\underline{w}^T \underline{\phi}(\underline{x})\right)$

$p(y=0) = 1 - \text{sigm}\left(\underline{w}^T \underline{\phi}(\underline{x})\right)$

or

$$p(y \mid \underline{x}, \underline{w}) = \left[\text{sigm}\left(\underline{w}^T \underline{\phi}\right)\right]^y \left[1 - \text{sigm}\left(\underline{w}^T \underline{\phi}\right)\right]^{1-y}$$

II. Likelihood of all observed data:

$$P(y_{1:n} \mid \underline{x}_{1:n}, \underline{w}) = \prod_{i=1}^{n} P(y_i \mid \underline{x}_i, \underline{w})$$

$$= \prod_{i=1}^{n} \left[\text{sigm}(\underline{w}^T \underline{\phi}(\underline{x}_i))\right]^{y_i} \left[1 - \text{sigm}(\underline{w}^T \underline{\phi}(\underline{x}_i))\right]^{1-y_i}$$

Then we obtain the best weight vector $\underline{w}$ using MLE:

$$\underline{w}^* = \max_{\underline{w}} \log P(y_{1:n} \mid \underline{x}_{1:n}, \underline{w})$$

$$= \max_{\underline{w}} \sum_{i=1}^{n} \left\{ y_i \, \text{sigm}(\underline{w}^T \underline{\phi}(\underline{x}_i)) + (1 - y_i)\left[1 - \text{sigm}(\underline{w}^T \underline{\phi}(\underline{x}_i))\right] \right\}$$

Finding max $\underline{w}$ is equivalent to minimizing the loss

$$L(\underline{w}) = -\sum_{i=1}^{n} \left\{ y_i \, \text{sigm}\left(\underline{w}^T \underline{\phi}(\underline{x}_i)\right) + (1-y_i)\left[1 - \text{sigm}\left(\underline{w}^T \underline{\phi}(\underline{x}_i)\right)\right] \right\}$$

this is called a **cross-entropy loss function.**
used also, eg, for DNNs that classify images.

III **Making Decisions** : say we have a point estimate $\underline{w}^*$

↳ $p(y \mid \underline{x}, \underline{w} = \underline{w}^*)$ is the prob. that $y=1$

↳ but you need to make a decision, $y=0$ or $y=1$

As before with decision making, need a **loss function**

$\ell(\hat{y}, y)$ = cost of picking $\hat{y}$ if the true value is y

choice of cost function is subjective.

For binary classification, $\ell(\hat{y}, y)$ is a 2×2 matrix

decision $\hat{y}$ \ y (true)	0	1
0	0	1.0
1	1.0	0

Given $\ell(\hat{y}, y)$, your decision is the $\hat{y}$ that minimizes expected loss

decision is $\min_{\hat{y}} \sum_{y=0,1} \ell(\hat{y}, y)\, p(y \mid \underline{x}, \underline{w} = \underline{w}^*)$

a different optimization problem for each $\underline{x}$

IV. Diagnostics for Classification

- split dataset into training & validation
- two important metrics are

1) accuracy $= \frac{1}{n_{samples}} \sum_{i=1}^{n_{samples}} 1(\hat{y}_i = y)$

balanced accuracy: average of recall (TP rate) for each class:

$\frac{TP}{TP+FN}$, $\frac{TN}{TN+FP}$ } avg

2) confusion matrix

actual \ predicted	0	1
0	TP	FN
1	FP	TN

would like the numbers on the diagonal to be large

V. Multi-class classification

observations $\underset{\sim}{x}_{1:n} = (\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n)$

targets $y_{1:n} = (y_1, \ldots, y_n)$ ← discrete labels

Now we have K possible values for labels so