Lecture 14, Mar 8, 2024

Logistic Regression

- Try to estimate $P[Y = i|\boldsymbol{x}, \boldsymbol{w}]$ where $\boldsymbol{y} \in C = \{1, 2, \dots, c\}$ are classes, $\boldsymbol{x}$ is a feature, and $\boldsymbol{w}$ are linear model weights
- Example: binary hypothesis ($c = 2$) with Bernoulli probabilities
  - Then $p(y = 1|x) = \dfrac{p(x|y = 1)p(y = 1)}{p(x|y = 1)p(y = 1) + p(x|y = 0)p(y = 0)} = \dfrac{1}{1 + \frac{p(x|y=0)p(y=0)}{p(x|y=1)p(y=1)}}$
  - We can write this as $\dfrac{1}{1 + e^{-\alpha}}$ where $\alpha = \log \dfrac{p(x|y = 0)p(y = 0)}{p(x|y = 1)p(y = 1)}$
- $\sigma(\alpha) = \dfrac{1}{1 + e^{-\alpha}}$ is the *sigmoid function*, which maps $\mathbb{R} \to (0, 1)$ which is useful for probabilities
  - Has an S shape with value of $\dfrac{1}{2}$ at 0
  - Note $\sigma(-\alpha) = 1 - \sigma(\alpha)$ and $\alpha = \log \dfrac{\sigma(\alpha)}{1 - \sigma(\alpha)}$
  - $\dfrac{\mathrm{d}\sigma}{\mathrm{d}\alpha} = \sigma(\alpha)(1 - \sigma(\alpha))$
  - We can classify $\hat{y} = 1$ if $\sigma(\alpha) > \dfrac{1}{2}$ or $\hat{y} = 0$ otherwise
- Our model is then $\hat{p}(y = 1|\boldsymbol{x}) = \dfrac{1}{1 + e^{-\boldsymbol{w}^T\boldsymbol{x}}} = \sigma(\boldsymbol{w}^T\boldsymbol{x})$, where we try to find the best weights $\boldsymbol{w}$
- Compared to Gaussian discriminant analysis, which has $2D$ for means and $D(D + 1)/2$ for covariances and priors, we only have D parameters and a lot less computation overall
- Consider a Bernoulli trial with $\theta = P[y = 1]$, so $P[y] = \theta^y(1 - \theta)^{1-y}$
  - Let $\theta = P[y = 1|\boldsymbol{x}, \boldsymbol{w}] = \sigma(\boldsymbol{w}^T\boldsymbol{x})$
  - For n trials, the NLL is $-\log \prod_{i=1}^{n} P[y_i|\boldsymbol{x}_i, \boldsymbol{w}] = -\sum_{i=1}^{n} \log(\theta_i^y(1 - \theta_i)^{1-y_i})) = -\sum_{i=1}^{n} y_i \log \theta_i + (1 - y_i) \log(1 - \theta_i)$ where $\theta_i = \sigma(\boldsymbol{w}^T\boldsymbol{x}_i) = \sigma_i$
  - This is the *cross-entropy* loss function
  - $\dfrac{\mathrm{d}}{\mathrm{d}w}\mathrm{NLL} = -\sum_{i=1}^{n}\left(y_i \dfrac{1}{\theta_i}\theta_i' + (1 - y_0)\dfrac{1}{1 - \theta_i}(-\theta_i')\right) = -\sum_{i=1}^{n}\left(y_i \dfrac{\theta_i'}{\theta_i} - (1 - y_0)\dfrac{\theta_i'}{1 - \theta_i}\right) = 0$
  - $\theta_i' = \sigma_i(1 - \sigma_i)\dfrac{\mathrm{d}}{\mathrm{d}w_j}\boldsymbol{w}^T\boldsymbol{x}_i = \sigma_i(1 - \sigma_i)x_{ij}$
  - $\dfrac{\theta_i'}{\theta_i} = (1 - \sigma_i)x_{ij} \implies \dfrac{\theta_i'}{1 - \theta_i'} = \sigma_i x_{ij}$
  - Therefore $\dfrac{\mathrm{d}}{\mathrm{d}w}\mathrm{NLL} = \sum_{i=1}^{n}(y_i - \sigma_i)x_{ij} = 0$
    - This can be interpret as the error multiplied by the observation
  - No closed-form solution; we can use methods such as gradient descent
    - The gradient vector is $\sum_{i=1}^{n}(y_i - \sigma_i)\boldsymbol{x}_i$
- Just like in linear regression, we're not restricted to just a single basis; we can change to e.g. a polynomial basis
  - Change of basis can make the space more linearly separable
  - Sometimes the problem is unsolvable as-is due to the data not being linearly separable
- For multiple classes, $p(c_k|\boldsymbol{x}, \boldsymbol{w}) = \dfrac{e^{\alpha_k}}{\sum_i e^{-\alpha_i}}$ where $\alpha_k = \boldsymbol{w}_k^T\boldsymbol{x}$
  - This is a softmax
  - This reduces to the same sigmoid function if we only have 2 classes
- We can also replace the sigmoid with tanh (and rescale to between 0 and 1)