13. Multiclass classification

- Adding label diversity
- Multiclass logistic regression
- One vs All, One vs One

binary: yes or no
+1, -1
heads, tails
success, fail

multiclass
dog, cat, mouse, horse
red, blue, green, yellow
~~1, 0, 1, 2~~

Some extensions are natural

Instead of binary labels $+1, -1$, just extend to K labels $1, 2, ..., K$

- Naive Bayes
- Decision trees
- K-nearest neighbors

$\Pr(y = \hat{y} \mid x)$

$\hat{y} \in \{1, \ldots, K\}$

In these schemes, all labels are treated equal

- $\Pr(y = 2)$ cannot be closer to $\Pr(y = 3)$ than it is to $\Pr(y = 100)$

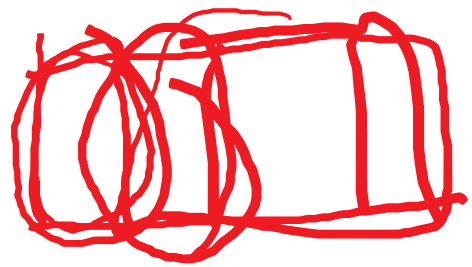
Multiclass as extension of binary classification

One vs All (OVA)

- Pick any binary classification scheme with soft output

$$g(x;\theta) = \begin{cases} \text{large and positive if} & \Pr(y=1) \text{ is high} \\ \text{small or negative if} & \Pr(y=1) \text{ is low} \end{cases}$$

- For each $k = 1, ..., K$, assign $y_i^{(k)} = \begin{cases} 1, & y_i = k \\ -1 & \text{else.} \end{cases}$

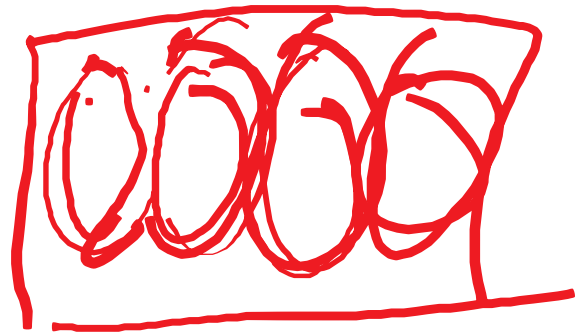


- Fit the binary classifier over $(x_i, y_i^{(k)})$, return classifier θ_k
- Then the multiclass classification scheme is a voting scheme

$$y = \underset{k=1,...,K}{\operatorname{argmax}} g(x;\theta_k)$$

- \+ Learns K classifiers independently (parallelizable)
- − Difficulty dealing with class imbalance

One vs one (OVO)



- Pick any binary classification scheme with soft output $g(x;\theta)$
- For each pair of classes j, k, pick out data with labels j or k
- Learn a binary classifier $\theta_{j,k}$ differentiating class j and k
- Multiclass classification scheme is again a voting scheme

$$y = \underset{k=1,\ldots,K}{\operatorname{argmax}} \sum_{i=1,\ldots,K} g(x;\theta_{i,k})$$

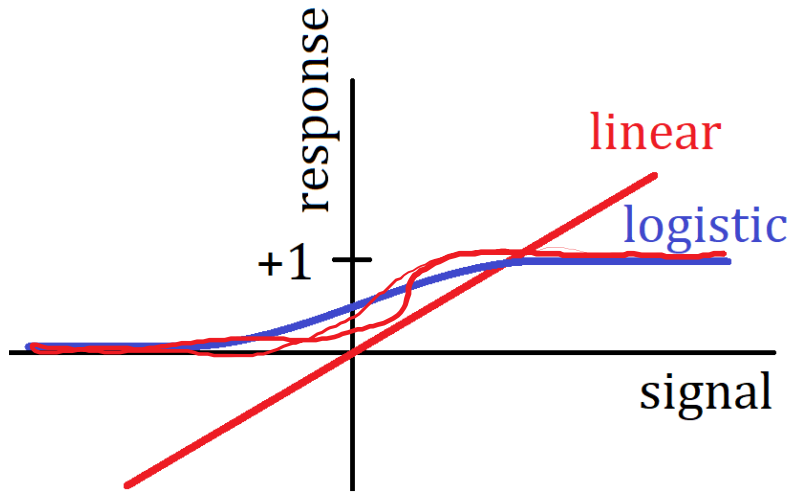
$\binom{K}{2}$

$\frac{K \cdot (K-1)}{2}$

- \+ better at dealing with imbalanced classes
- – each classifier may have very little data to work with
- – requires learning $O(K^2)$ classifiers (but parallelizable)

Multiclass Logistic regression

Remember logistic regression



- **Linear**: response is linear $(y = x^T\theta)$
- **Logistic**: logit is linear $\left(\log\left(\frac{\Pr(y=1)}{\Pr(y=0)}\right) = x^T\theta\right)$

  Interpretation: Tapering out signals that are too strong

Derivation: logistic regression

- Define $p = \Pr(y = 1)$. Then

$$\log\left(\frac{p}{1-p}\right) = x^T\theta$$

- Solve for p:

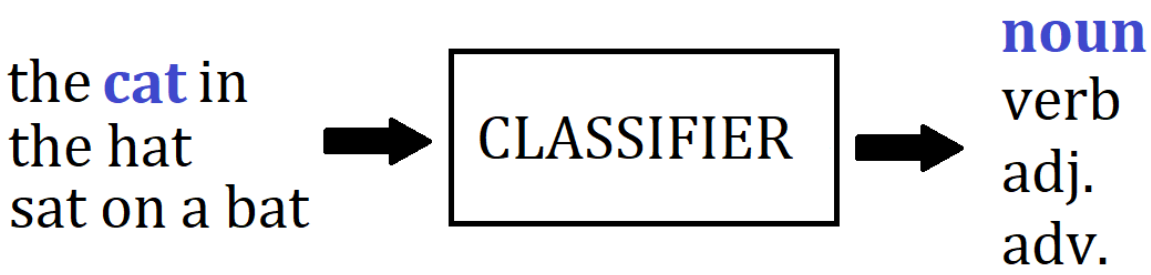
$$p = \frac{e^{\theta^T x}}{1 + e^{\theta^T x}} = \sigma(\theta^T x)$$

- Log likelihood:

$$\log(\Pr(y)) = \begin{cases} \log(\sigma(\theta^T x)) & \text{if } y = 1 \\ \log(1 - \sigma(\theta^T x)) & \text{if } y = 0 \end{cases}$$

$$= \begin{cases} \log(\sigma(\theta^T x)) & \text{if } \hat{y} = 1 \\ \log(-\sigma(\theta^T x)) & \text{if } \hat{y} = -1 \end{cases}$$

Extend to multiclass logistic regression



K classes

(parts of speech, type of image, diagnosis, recommendation,...)

Logistic assumption

$$\log\left(\frac{\Pr(Y=k)}{\Pr(Y=j)}\right) = \theta_k^T X - \theta_j^T X, \quad \theta_k \in \mathbb{R}^n \quad k = 1, ..., K$$

$j = 1, ..., K$

Derivation: multiclass logistic regression

- Define $p_k = \Pr(Y = k)$, $k = 1, ..., L$
- Logistic assumption

$$\log\left(\frac{p_k}{p_j}\right) = \theta_k^T x - \theta_j^T x, \quad \theta_k \in \mathbb{R}^n, \ k = 1, ..., K$$

- True if for some unknown $c > 0$

$$\log(cp_k) = \theta_k^T x \iff p_k = c^{-1} \exp(\theta_k^T x).$$

- Since $\sum_i \Pr(Y = i) = 1$,

$$1 = \sum_{i=1}^{L} c^{-1} \exp(\theta_i^T x) \iff c = \sum_{i=1}^{L} \exp(\theta_i^T x)$$

- Therefore,

$$\Pr(Y = k) = \frac{\exp(\theta_k^T x)}{\sum_{i=1}^{L} \exp(\theta_i^T x)}$$

Softmax function

- Input space: probability simplex

$$\Delta_{n-1} = \left\{ p \in \mathbb{R}^n : p[k] \geq 0 \; \forall k, \;\; \sum_{i=1}^{n} p[i] = 1 \right\}$$

- Max function: $f_{\max}(p) = \max_{i=1,\ldots,n} p[i]$
- Softmax function ($\mu > 0$):

$$f_\mu(p) = \frac{1}{\mu} \log \left(\frac{1}{L} \sum_{i=1}^{L} \exp\left(\mu \, p[i]\right) \right)$$

- Limits

$$\lim_{\mu \to \infty} f_\mu(p) = f_{\max}(p), \qquad \lim_{\mu \to 0} f_\mu(p) = f_{\text{mean}}(p)$$

Softmax weights

Softmax function

$$f_\mu(p) = \frac{1}{\mu} \log \left(\frac{1}{L} \sum_{i=1}^{L} \exp\left(\mu\, p[i]\right) \right)$$

Softmax weights

$$\frac{\partial f_\mu}{\partial p[k]}(p) = \frac{\frac{1}{\mu}}{\frac{1}{L}\sum_{i=1}^{L} \exp(\mu p_i)} \cdot \frac{1}{L} \exp(\mu p_k)$$

$$= \frac{\exp(\mu p_k)}{\sum_{i=1}^{L} \exp(\mu p_i)}$$

Softmax weights

Softmax function

$$f_\mu(p) = \frac{1}{\mu} \log \left(\frac{1}{L} \sum_{i=1}^{L} \exp\left(\mu\, p[i]\right) \right)$$

Softmax weights

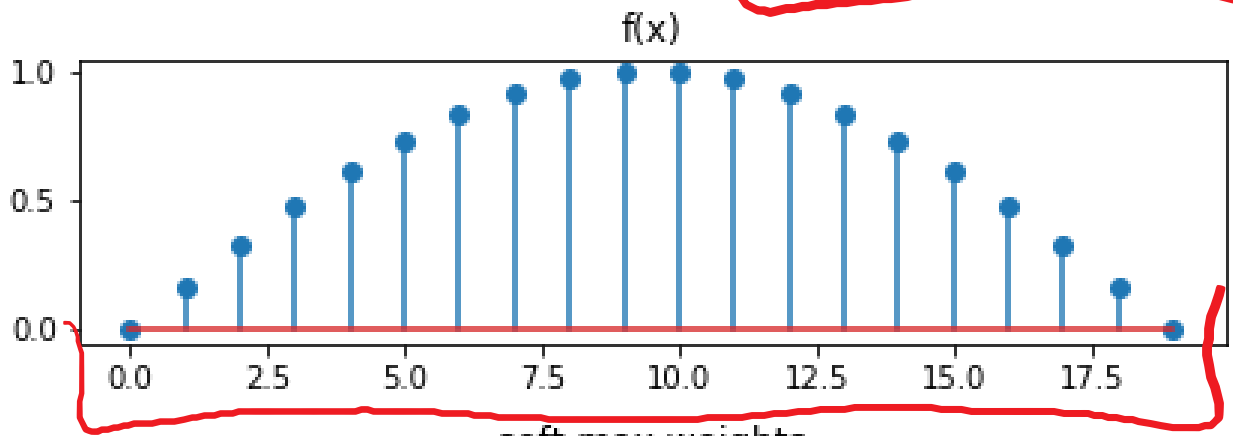
$$\frac{\partial f_\mu}{\partial p[k]}(p) = \frac{\exp(\mu p[k])}{\sum_{i=1}^{L} \exp(\mu p[i])}$$

$\|\theta_t\|_2$

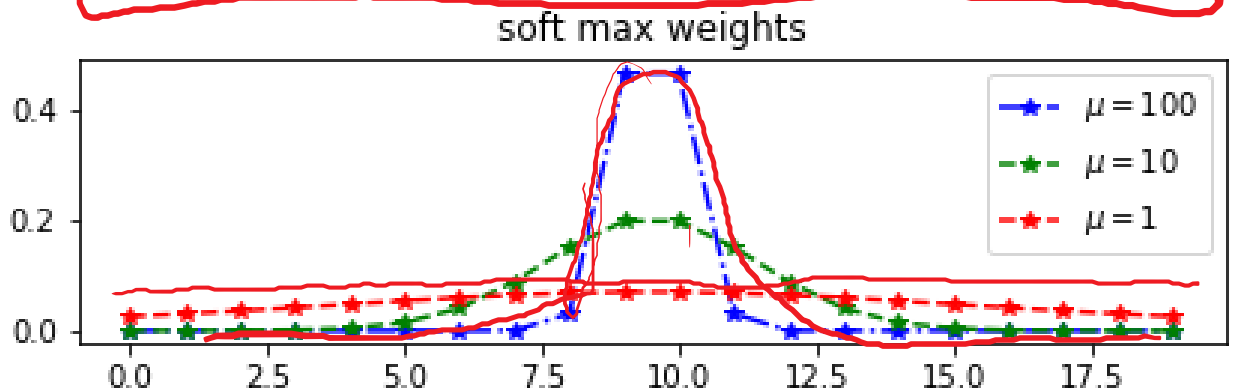
Softmax

$$f_\mu(p) = \frac{1}{\mu} \log \left(\frac{1}{L} \sum_{i=1}^{L} \exp\left(\mu\, p[i]\right) \right),$$

$$\frac{\partial f_\mu}{\partial p[k]}(p) = \frac{\exp(\mu p[k])}{\sum_{i=1}^{L} \exp(\mu p[i])}$$



f_μ

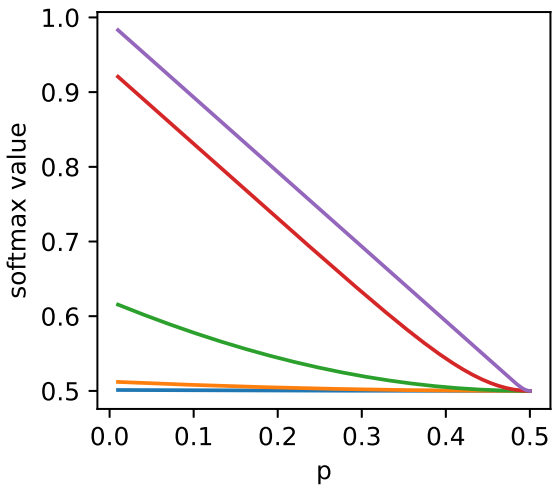


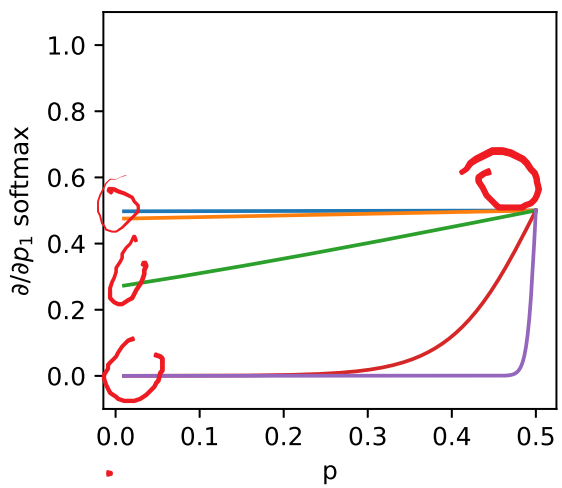
Softmax, binary case

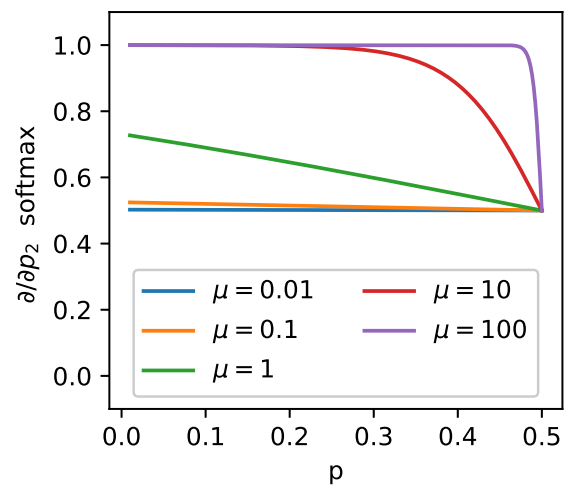
$$p_k = \frac{e^{\theta_k^T x}}{\sum_i e^{\theta_i^T x}}$$

$$p = \Pr(Y = 1), \qquad f_\mu(p) = \frac{1}{\mu} \log\left(\frac{\exp(\mu\, p) + \exp(\mu\,(1-p))}{2}\right)$$

$$\frac{\partial f_\mu}{\partial p}(p) = \frac{\exp(\mu p)}{\exp(\mu p) + \exp(\mu(1-p))} = \sigma(\mu(p-(1-p)))$$







Multiclass logistic regression

Since $\sum_k \Pr(Y = k) = 1$, it must be that

$$\Pr(Y = k) = \frac{\exp(\theta_k^T X)}{\sum_{j=1}^K \exp(\theta_j^T X)}$$

Log likelihood of (x_i, y_i):

$$\sum_{i=1}^m \log(\Pr(Y = y_i)) = \log \sum_{i=1}^m \exp(\theta_{y_i}^T x_i) - \log \sum_{i=1}^m \sum_{j=1}^K \exp(\theta_j^T X)$$

independent of labels

Multiclass logistic regression

$$\underset{\theta \in \mathbb{R}^{n \times k}}{\text{maximize}} \quad \log \sum_{i=1}^m \exp(\theta_{y_i}^T x_i) - \sum_{i=1}^m \log \sum_{j=1}^K \exp(\theta_j^T X)$$

Connection to distribution divergence

Define the label distribution as

$$p_k(X) := \begin{cases} 1 & \text{if } Y = k \\ 0 & \text{else.} \end{cases}$$

and the predicted distribution as

$$q_k(X) := \frac{\exp(\theta_k^T X)}{\sum_{j=1}^{K} \exp(\theta_j^T X)}$$

The Kullback Leibler divergence between distributions p, q

$$D_{KL}(p||q) := \underbrace{\sum_{x \in \mathcal{X}} p(x) \log(p(x))}_{\text{label entropy}} - \underbrace{\sum_{x \in \mathcal{X}} p(x) \log(q(x))}_{=:H(p,q),\ \text{cross entropy}}$$

and the cross-entropy can be further expanded as

$$H(p,q) = \log \sum_{i=1}^{m} \exp(\theta_{y_i}^T x_i) - \log \sum_{i=1}^{m} \sum_{j=1}^{K} \exp(\theta_j^T x_i) = \text{log likelihood}$$