Logistic Regression

Logisitc regression defines the conditional probability directly and is a **discriminative model** rather than a generative model.

- Start with the scoring function $\boldsymbol{\theta} \cdot$ t measures the comptability between the features

- To make sure it
  $\exp \boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{x},$

- We then normalize it by dividing it over all po
  $y \in \mathcal{Y}$ and get a probability:

$$p(y|\boldsymbol{x};\boldsymbol{\theta}) = \frac{\exp \boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{x}, y)}{\sum_{y' \in \mathcal{Y}} \exp \boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{x}, y')}$$

Logistic Regression

The weights are estimat **kelihood**.
Give a data set D $_{i=1}$ itional
likelihood is:

$$\log p(y^{(1:N)} | \boldsymbol{x}^{(1:N)}; \theta) = \sum^{N} \log p(y^{(i)} | \boldsymbol{x}$$

$$\sum_{i=1} \quad \mathcal{Y} \quad \exp \boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{x}^{(i)}, y')$$

Or they can be estimated by minimizing the logis
(or cross-entropy loss):

$$\mathcal{L}_{\text{LOGREG}} = -\sum_{i=1}^{N} \left(\boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{x}^{(i)}, y^{(i)}) - \log \sum_{y' \in \mathcal{Y}} \exp \left(\boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{x}^{(i)}, y') \right) \right)$$

Logistic Regression Objective

Loss of a single sample:

$$\ell_{\text{LOGREG}} = \quad \cdots \quad \cdot \boldsymbol{f}(\boldsymbol{x}^{(i)}, y')\Big)$$

Gradient of Logistic Regression

The gradient with respe

$$
\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\theta}} &= -\boldsymbol{f}(\boldsymbol{x}^{(i)}, y^{(i)}) + \frac{1}{\sum_{y' \in \mathcal{Y}} \exp(\boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{x}^{(i)}, y'))} \\
&\quad \times \sum_{y' \in \mathcal{Y}} \exp(\boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{x}^{(i)}, y')) \times \boldsymbol{f}(\boldsymbol{x}^{(i)}, y') \\
&= -\boldsymbol{f}(\boldsymbol{x}^{(i)}, y^{(i)}) + \sum_{y' \in \mathcal{Y}} \frac{\exp(\boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{x}^{(i)}, y'))}{\sum_{y'' \in \mathcal{Y}} \exp(\boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{x}^{(i)}, y''))} \times \boldsymbol{f}(\boldsymbol{x}^{(i)}, y') \\
&= -\boldsymbol{f}(\boldsymbol{x}^{(i)}, y^{(i)}) + \sum_{y' \in \mathcal{Y}} P(y' | \boldsymbol{x}^{(i)}; \boldsymbol{\theta}) \times \boldsymbol{f}(\boldsymbol{x}^{(i)}, y') \\
&= -\boldsymbol{f}(\boldsymbol{x}^{(i)}, y^{(i)}) + E_{Y|X}[\boldsymbol{f}(\boldsymbol{x}^{(i)}, y)]
\end{aligned}
$$

Application of the chain rule in calculus, expectation

Gradient of Logistic Regression

$$\frac{\partial}{\partial \boldsymbol{\theta}} = -\boldsymbol{f}(\boldsymbol{x}^{(i)}, y^{(i)}) + \sum_{y' \in \mathcal{Y}} P(y' | \boldsymbol{x}^{(i)}; \boldsymbol{\theta}) \times \boldsymbol{f}(\boldsymbol{x}^{(i)}, y')$$

$$= -\boldsymbol{f}(\boldsymbol{x}^{(i)}, y^{(i)}) + E_{Y|X}[\boldsymbol{f}(\boldsymbol{x}$$

- This is a very nic

  - The gradient equals to the difference b
feature counts under the current mod $(\boldsymbol{x}^{(i)}, y)]$ and the observed feature counts $\boldsymbol{f}$

  - The loss is minimized if the feature counts under the current model and the observed feature counts are the same
- The power of Logistic Regression model is that you can use arbitrary number of features without making any independence assumptions. The allows creative feature engineering to improve the performance of the model.

Digression: Expectation

Expectation is the mean
discrete case. Let X be a random variable:

$$\mathbb{E}[X] = \sum_{x \in X} xP(x) =$$

$$[g(X)] = \quad g(x)p(x)$$

$$)$$

$$x,y$$

Let X be the random variable which is the value of rolling a single dice:

$$\mathbb{E}[x] = \sum_{x=1}^{6} xP(y) = \frac{1}{6}\sum_{x=1}^{6} x = \frac{21}{6} = 3.5$$

Digression: Linearity of Expectations

$$\mathbb{E}[X + Y]$$

$$= \sum_{x,y} P(x, y)x + \sum_{x,y} P(x, y)y$$

$$(x, y)\Big)\, y$$

$$= \mathbb{E}[X] + \mathbb{E}[Y]$$

Let the Y be the random variable for the sum of two dice rolled.
Expected value of Y:

$$\mathbb{E}[Y] = \mathbb{E}[X] + \mathbb{E}[X]$$

When two random variables are independent:

$$\mathbb{E}[X, Y] = \mathbb{E}[X] \times \mathbb{E}[X]$$

Digression: Variance

- Variance of a rand the random variabl s or whether they vary a lot

$$Var(X) = \mathbb{E}[(X - \mathbb{E}[X])^2]$$

$$
\begin{aligned}
\mathbb{E}[(X - \mathbb{E}[X & \ldots \sum_x \\
&= \sum_x (x^2 - 2\mathbb{E}[X]x \ldots \\
&= \sum_x x^2 P(x) - 2\mathbb{E}[X] \sum_x xP(x) + (\mathbb{E}[X])^2 \sum_x P(x) \\
&= \mathbb{E}[X^2] - 2(\mathbb{E}[X])^2 + (\mathbb{E}[X])^2 \\
&= \mathbb{E}[X^2] - (\mathbb{E}[X])^2
\end{aligned}
$$

- σ^2 denotes the variance, and σ is the standard deviation.

Regularization

- $L2$ regularization: add a multiple of the squared norm $\frac{\lambda}{2}\|\boldsymbol{\theta}\|_2^2$ to the minimization objective

- Regularization forces the estimator to trad

on the training data against the norm of the weights, and thus helps relieve o

The overall loss of a tra

$$\mathcal{L}_{\text{LOGREG}} = \frac{\lambda}{2}\|\boldsymbol{\theta}\|_2^2 - \sum_{i=1}^{N}\left(\boldsymbol{\theta}\cdot\boldsymbol{f}(\boldsymbol{x}^{(i)}, y^{(i)}) - \sum_{y'\in\mathcal{Y}}\left(\boldsymbol{\theta}\cdot\boldsymbol{f}(\boldsymbol{x}^{(i)}, y)\right)\right)$$

Regularized gradient

Derivative of the L_2 norm:

$$\frac{\lambda}{2}\|\boldsymbol{\theta}\|_2^2 = \frac{\lambda}{2}\left(\left(\sum_{j=1}^{N}\theta_j^2\right)^{\frac{1}{2}}\right)^2 = \frac{\lambda}{2}\sum_{j=1}^{N}\theta_j^2$$

$$\frac{\partial}{\partial\theta_k}\frac{\lambda}{2}\|\boldsymbol{\theta}\|_2^2 = \frac{\partial}{\partial\theta_k}\frac{\lambda}{2}\sum_{j=1}^{N}\theta_j^2 = \sum_{j=1}^{N}\frac{\lambda}{2}\frac{\partial}{\partial\theta_k}\theta_j^2 = \lambda\theta_k$$

$$\frac{\partial}{\partial\theta_k}\theta_j^2 = \begin{cases} 2\theta_k & j = k \\ 0 & \text{otherwise} \end{cases}$$

Gradient of regularized loss:

$$\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\text{LOGREG}} = \lambda\boldsymbol{\theta} - \sum_{i=1}^{N}\left(\boldsymbol{f}(\boldsymbol{x}^{(i)}, y^{(i)}) - E_{Y|X}[\boldsymbol{f}(\boldsymbol{x}^{(i)}, y')]\right)$$

Batch Optimization vs online optimization

Gradient Descent vs Stochastic Gradient Descent. In **batch optimization**, each update to the weights is based o

dataset. One such algorithm is **gradien** ich iteratively updates the weight

$$\boldsymbol{\theta}^{(t+1)} \leftarrow \boldsymbol{\theta}^{(t)}$$

where $\nabla_{\boldsymbol{\theta}}\mathcal{L}$ is the gradient computed over the entire training set, $\eta^{(t)}$ is the **learning rate** at iteration t.

Variations of Gradient Descent

- **Online learning** algorithms make upd h the data. In **stochasttic gradient des** e approximate gradient is co update is mad
- In **mini-batch** stochastic gradien nt is computed over a small subset of instance

Generalized gradient descent algorithm

The function BATCHER [illegible] B batches such that each instance appears in exactly one batch. In stochastic gradient descent, $B = N$; in gradient descent, $B = 1$; in mini-batch stochastic gradient descent, $1 < B < N$.

procedure GRADIENTDESCENT($x^{1:N}$, [illegible], BATCHER, T_{MAX})
  $t \leftarrow 0$
  $\boldsymbol{\theta}^{(0)} \leftarrow \mathbf{0}$
  repeat
    $(\boldsymbol{b}^{(1)}, \boldsymbol{b}^{(2)}, \cdots, \boldsymbol{b}^{(B)}) \leftarrow$ BA[illegible]
    for $n \in \{1, 2, \cdots, B\}$ **do**
      $t \leftarrow t + 1$
      $\boldsymbol{\theta}^{(t)} \leftarrow \boldsymbol{\theta}^{(t-1)} - \eta^{(t)} \nabla \mathcal{L}(\boldsymbol{\theta}^{(t-1)}; \boldsymbol{x}^{(b_1^{(n)}, b_2^{(n)}, \cdots)}, \boldsymbol{y}^{(b_1^{(n)}, b_2^{(n)}, \cdots)})$
      if Converged($\boldsymbol{\theta}^{(1,2,\cdots,t)}$) **then return** $\boldsymbol{\theta}^{(t)}$
      end if
    end for
  until $t = T_{MAX}$
end procedure

Binary logistic regression is a special case of multinominal logistic regression

$$
\begin{aligned}
P(y=1|\boldsymbol{x}) &= \frac{\exp(\sum_k \theta_k f_k(y=1,\boldsymbol{x}))}{\exp(\sum_k \theta_k f_k(y=1,\boldsymbol{x})) + \exp(\sum_k \theta_k f_k(y=0,\boldsymbol{x}))} \\
&= \frac{\exp(\sum_k \theta_k f_k(y=1,\boldsymbol{x}))}{\exp(\sum_k \theta_k f_k(y=1,\boldsymbol{x})) + \exp(\sum_k \theta_k f_k(y=0,\boldsymbol{x}))} \\
&\quad \times \frac{\exp(-\sum_k \theta_k f_k(y=1,\boldsymbol{x}))}{\exp(-\sum_k \theta_k f_k(y=1,\boldsymbol{x}))} \\
&= \frac{1}{1+\exp(\sum_k \theta_k f_k(y=0,\boldsymbol{x}) - \sum_k \theta_k f_k(y=1,\boldsymbol{x}))} \\
&= \frac{1}{1+\exp(\sum_k \theta_k (f_k(y=0,\boldsymbol{x}) - f_k(y=1,\boldsymbol{x})))} \\
&= \frac{1}{1+\exp(\sum_k -\theta_k f'_k(\boldsymbol{x}))} = \sigma(\boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{x}))
\end{aligned}
$$

Note: For binary classification, you only need to pay attention to the positive class.

Logistic Regression: features and weights

	not	funny	painful	ok	overall	story	good	jokes	*bias*
POS	f_1	f_4	f_7	f_{10}	f_{13}			22	f_{25}
NEG	f_2	f_5	f_8	f_{11}	f_{14}			23	f_{26}
NEU	f_3	f_6					1	f_{24}	f_{27}

POS	-1	2	-2.5	0.5	0.2	.08	1.5	0.8	1.2
NEG	2	-2	1.8	-0.5	0.1	-0.6	-2	-1.2	0.8
NEU	-0.4	-0.9	-1.5	2	1	-0.2	-1.2	-0.3	0.4

e.g., $f_1(x, y) = 1$ if x= "not" $\wedge$ y= "POS", $\theta_1 = -1$

Note: The features $\boldsymbol{f}(\boldsymbol{x}, y)$ and $\boldsymbol{\theta}$ are presented in a table rather than a vector due to limitation of space. Mathematically they should still be viewed as vectors

Logistic Regression: Inference

$\boldsymbol{f}(\boldsymbol{x}, y)$

	not	funny	painful	ok	overall	story	good	jokes	*bias*
POS	f_1	f_4	f_7	f_{10}	f_{13}			22	f_{25}
NEG	f_2	f_5	f_8	f_{11}	f_{14}			23	f_{26}
NEU	f_3	f_6					1	f_{24}	f_{27}

POS	-1	2	-2.5	0.5	0.2	.08	1.5	0.8	1.2
NEG	2	-2	1.8	-0.5	0.1	-0.6	-2	-1.2	0.8
NEU	-0.4	-0.9	-1.5	2	1	-0.2	-1.2	-0.3	0.4

Test instance: "funny, smart, and visually stunning"

$$p(y|\boldsymbol{x}) = \frac{\exp(\boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{x}, y))}{\sum_y' \exp(\boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{x}, y'))} = \frac{\exp(\sum_k \theta_k f_k(\boldsymbol{x}, y))}{\sum_{y'} \exp(\sum_k \theta_k f_k(\boldsymbol{x}, y'))}$$

Logistic Regression: Inference

	not	funny	painful	ok	overall	story	good	jokes	*bias*
POS	f_1	f_4	f_7	f_{10}	f_{13}	f_{16}	f_{19}	f_{22}	f_{25}
NEG	f_2	f_5	f_8	f_{11}	f_{14}	[illegible]	[illegible]	f_{23}	f_{26}
NEU	f_3	f_6	f_9	f_{12}	f_{15}	[illegible]	[illegible]	f_{24}	f_{27}

POS	-1	2	-2.5	0.5	0.2	0.8	1.5	0.8	1.2
NEG	2	-2	1.8	-0.5	0.1	-0.6	-2	-1.2	0.8
NEU	-0.4	-0.9	-1.5	2	1	-0.2	-1.2	-0.3	0.4

Test instance: “funny, smart, and visually stunning”

$$p(y = POS|\boldsymbol{x})$$

$$= \frac{\exp(\theta_4 f_4 + \theta_{25} f_{25})}{\exp(\theta_4 f_4 + \theta_{25} f_{25}) + \exp(\theta_5 f_5 + \theta_{26} f_{26}) + \exp(\theta_6 f_6 + \theta_{27} f_{27})}$$

$$= \frac{\exp(2 + 1.2)}{\exp(2 + 1.2) + \exp(-2 + 0.8) + \exp(-0.9 + 0.4)} = 0.9643$$

Logistic Regression: Inference

	not	funny	painful	ok	overall	story	good	jokes	*bias*
POS	f_1	f_4	f_7	f_{10}	f_{13}	f_{16}	f_{19}	f_{22}	f_{25}
NEG	f_2	f_5	f_8	f_{11}	f_{14}			23	f_{26}
NEU	f_3	f_6	f_9	f_{12}	f_{15}			24	f_{27}

POS	-1	2	-2.5	0.5	0.2	0.8	1.5	0.8	1.2
NEG	2	-2	1.8	-0.5	0.1	-0.6	-2	-1.2	0.8
NEU	-0.4	-0.9	-1.5	2	1	-0.2	-1.2	-0.3	0.4

Test instance: "funny, smart, and visually stunning"

$$p(y = NEG|\boldsymbol{x})$$

$$= \frac{\exp(\theta_5 f_5 + \theta_{26} f_{26})}{\exp(\theta_4 f_4 + \theta_{25} f_{25}) + \exp(\theta_5 f_5 + \theta_{26} f_{26}) + \exp(\theta_6 f_6 + \theta_{27} f_{27})}$$

$$= \frac{\exp(-2 + 0.8)}{\exp(2 + 1.2) + \exp(-2 + 0.8) + \exp(-0.9 + 0.4)} = 0.0118$$

Logistic Regression: Inference

	not	funny	painful	ok	overall	story	good	jokes	*bias*
POS	f_1	f_4	f_7	f_{10}	f_{13}	f_{16}	f_{19}	f_{22}	f_{25}
NEG	f_2	f_5	f_8	f_{11}	f_{14}	[illegible]	[illegible]	f_{23}	f_{26}
NEU	f_3	f_6	f_9	f_{12}	f_{15}	[illegible]	[illegible]	f_{24}	f_{27}

POS	-1	2	-2.5	0.5	0.2	-0.8	1.5	0.8	1.2
NEG	2	-2	1.8	-0.5	0.1	-0.6	-2	-1.2	0.8
NEU	-0.4	-0.9	-1.5	2	1	-0.2	-1.2	-0.3	0.4

Test instance: "funny, smart, and visually stunning"

$$p(y = NEU|\boldsymbol{x})$$

$$= \frac{\exp(\theta_6 f_6 + \theta_{27} f_{27})}{\exp(\theta_4 f_4 + \theta_{25} f_{25}) + \exp(\theta_5 f_5 + \theta_{26} f_{26}) + \exp(\theta_6 f_6 + \theta_{27} f_{27})}$$

$$= \frac{\exp(-9 + 0.4)}{\exp(2 + 1.2) + \exp(-2 + 0.8) + \exp(-0.9 + 0.4)} = 0.0238$$

Logistic Regression: Parameter estimation

,

	not	funny	painful	ok	overall	story	good	jokes	*bias*
POS	f_1	f_4	f_7	f_{10}	f_{13}			22	f_{25}
NEG	f_2	f_5	f_8	f_{11}	f_{14}			23	f_{26}
NEU	f_3	f_6					1	f_{24}	f_{27}

POS	-1	2	-2.5	0.5	0.2	.08	1.5	0.8	1.2
NEG	2	-2	1.8	-0.5	0.1	-0.6	-2	-1.2	0.8
NEU	-0.4	-0.9	-1.5	2	1	-0.2	-1.2	-0.3	0.4

Training instance: y = NEG, x = "not funny at all"

$$\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\text{LOGREG}} = -\sum_{i=1}^{N} \boldsymbol{f}(\boldsymbol{x}^{(i)}, y^{(i)}) + \sum_{i=1}^{N} \sum_{y' \in Y} p(y' | \boldsymbol{x}^{(i)}) \boldsymbol{f}(\boldsymbol{x}^{(i)}, y')$$

Logistic Regression: parameter estimation

	not	funny	painful	ok	overall	story	good	jokes	*bias*
POS	f_1	f_4	f_7	f_{10}	f_{13}			22	f_{25}
NEG	f_2	f_5	f_8	f_{11}	f_{14}			23	f_{26}
NEU	f_3	f_6	f_9	f_{12}	f_{15}	18	21	24	f_{27}

POS	-1	2	-2.5	0.5	0.2	.08	1.5	0.8	1.2
NEG	2	-2	1.8	-0.5	0.1	-0.6	-2	-1.2	0.8
NEU	-0.4	-0.9	-1.5	2	1	-0.2	-1.2	-0.3	0.4

Training instance: y = NEG, x = "not funny at all"

$$p(y = NEG|x)$$

$$= \frac{\exp(\theta_2 f_2 + \theta_5 f_5 + \theta_{26} f_{26})}{\exp(\theta_1 f_1 + \theta_4 f_4 + \theta_{25} f_{25})) + \exp(\theta_2 f_2 + \theta_5 f_5 + \theta_{26} f_{26})) + \exp(\theta_3 f_3 + \theta_6 f_6 + \theta_{27} f_{27}))}$$

$$= \frac{\exp(2 - 2 + 0.8)}{\exp(2 - 1 + 1.2) + \exp(2 - 2 + 0.8) + \exp(-0.9 - 0.4 + 0.4)} = 0.1909$$

Logistic Regression: parameter estimation

	not	funny	painful	ok	overall	story	good	jokes	*bias*
POS	f_1	f_4	f_7	f_{10}	f_{13}			22	f_{25}
NEG	f_2	f_5	f_8	f_{11}	f_{14}			23	f_{26}
NEU	f_3	f_6	f_9	f_{12}	f_{15}	18	21	24	f_{27}

POS	-1	2	-2.5	0.5	0.2	.08	1.5	0.8	1.2
NEG	2	-2	1.8	-0.5	0.1	-0.6	-2	-1.2	0.8
NEU	-0.4	-0.9	-1.5	2	1	-0.2	-1.2	-0.3	0.4

Training instance: y = NEG, x = “not funny at all”

$$p(y = POS|x)$$

$$= \frac{\exp(\theta_1 f_1 + \theta_4 f_4 + \theta_{25} f_{25}))}{\exp(\theta_1 f_1 + \theta_4 f_4 + \theta_{25} f_{25})) + \exp(\theta_2 f_2 + \theta_5 f_5 + \theta_{26} f_{26})) + \exp(\theta_3 f_3 + \theta_6 f_6 + \theta_{27} f_{27}))}$$

$$= \frac{\exp(2 - 1 + 1.2)}{\exp(2 - 1 + 1.2) + \exp(2 - 2 + 0.8) + \exp(-0.9 - 0.4 + 0.4)} = 0.7742$$

Logistic Regression: Parameter estimation

	not	funny	painful	ok	overall	story	good	jokes	*bias*
POS	f_1	f_4	f_7	f_{10}	f_{13}			22	f_{25}
NEG	f_2	f_5	f_8	f_{11}	f_{14}			23	f_{26}
NEU	f_3	f_6	f_9	f_{12}	f_{15}	18	21	24	f_{27}

POS	-1	2	-2.5	0.5	0.2	.08	1.5	0.8	1.2
NEG	2	-2	1.8	-0.5	0.1	-0.6	-2	-1.2	0.8
NEU	-0.4	-0.9	-1.5	2	1	-0.2	-1.2	-0.3	0.4

Training instance: y = NEG, x = "not funny at all"

$$
\begin{aligned}
& p(y = NEU|x) \\
&= \frac{\exp(\theta_3 f_3 + \theta_6 f_6 + \theta_{27} f_{27}))}{\exp(\theta_1 f_1 + \theta_4 f_4 + \theta_{25} f_{25})) + \exp(\theta_2 f_2 + \theta_5 f_5 + \theta_{26} f_{26})) + \exp(\theta_3 f_3 + \theta_6 f_6 + \theta_{27} f_{27}))} \\
&= \frac{\exp(-0.9 - 4 + 0.4)}{\exp(2 - 1 + 1.2) + \exp(2 - 2 + 0.8) + \exp(-0.9 - 0.4 + 0.4)} = 0.0349
\end{aligned}
$$

Logistic Regression: Parameter estimation

	not	funny	painful	ok	overall	story	good	jokes	*bias*
POS	f_1	f_4	f_7	f_{10}	f_{13}	f_{16}	f_{19}	f_{22}	f_{25}
NEG	f_2	f_5	f_8	f_{11}	f_{14}			f_{23}	f_{26}
NEU	f_3	f_6	f_9	f_{12}	f_{15}			f_{24}	f_{27}

POS	-1	2	-2.5	0.5	0.2	0.8	1.5	0.8	1.2
NEG	2	-2	1.8	-0.5	0.1	-0.6	-2	-1.2	0.8
NEU	-0.4	-0.9	-1.5	2	1	-0.2	-1.2	-0.3	0.4

Training instance: y = NEG, x = "not funny at all"

$$gradient[\theta_1] = -0 + 0.7742, gradient[\theta_4] = -0 + 0.7742$$
$$gradient[\theta_2] = -1 + 0.1909, gradient[\theta_5] = -1 + 0.1909$$
$$gradient[\theta_3] = -0 + 0.0349, gradient[\theta_6] = -0 + 0.0349$$

Logistic Regression: Parameter estimation

	not	funny	painful	ok	overall	story	good	jokes	*bias*
POS	f_1	f_4	f_7	f_{10}	f_{13}	f_{16}	f_{19}	f_{22}	f_{25}
NEG	f_2	f_5	f_8	f_{11}	f_{14}			f_{23}	f_{26}
NEU	f_3	f_6	f_9	f_{12}	f_{15}			f_{24}	f_{27}

POS	-1	2	-2.5	0.5	0.2	.08	1.5	0.8	1.2
NEG	2	-2	1.8	-0.5	0.1	-0.6	-2	-1.2	0.8
NEU	-0.4	-0.9	-1.5	2	1	-0.2	-1.2	-0.3	0.4

Training instance: y = NEG, x = “not funny at all”

$$gradient[\theta_{25}] = -0 + 0.7742$$
$$gradient[\theta_{26}] = -1 + 0.1909$$
$$gradient[\theta_{27}] = -0 + 0.0349$$

Note: The “bias” is the feature that always fires.

Some observations about the gradient for Logistic Regression

- The gradient of all features that predict the sa updated by the same amount.
- A feature is pen proportion

Discussion question:

- Assuming that the weights $\boldsymbol{\theta}$ are initialized as , how would the first iteration of the Logistic Regression Training proceed?

Adding a regularization term

POS	-1	2	-2.5	0.5	0.2	.08	1.5	0.8	1.2
NEG	2	-2	1.8	-0.5	0.1	-0.6	-2	-1.2	0.8
NEU	-0.4	-0.9	-1.5	2	1	-0.2	-1.2	-0.3	0.4

Training instance: y = NEG, x = "not funny at all"

$$\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\text{LOGREG}} = \;[illegible]_{i=1}\;[illegible]\,(y'|\boldsymbol{x}^{(i)})\boldsymbol{f}(\boldsymbol{x}^{(i)}, y')$$

Let $\lambda = 0.5$:

gradient$[\theta_1] = -0.5 - 0 + 0.7742$, gradient$[\theta_4] = 1 - 0 + 0.7742$

gradient$[\theta_2] = 1 - 1 + 0.1909$, gradient$[\theta_5] = -1 - 1 + 0.1909$

gradient$[\theta_3] = -0.2 - 0 + 0.0349$, gradient$[\theta_6] = -0.45 - 0 + 0.0349$