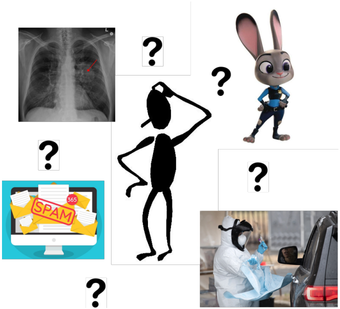
2. Logistic regression

- Binary classification
- Logistic regression
- This is a broad overview, we will revisit each section in depth

Binary classification

Mind-reading task

- Will Alice like the movie?
- Will Bob click on this link?
- Is this cancer?
- Is this spam?
- Is this a terrorist?



Features: (age, height, political leanings, historical decisions, friends)

Labels: (yes or no)

Mind-reading task

Features: (age, height, political leanings, historical decisions, friends)

Labels: (yes or no)

Mind-reading task

Features: (age, height, political leanings, **historical decisions**, friends)

Labels: (yes or no)

- Alice liked Finding Nemo, so she'll like Zootopia

Mind-reading task

Features: (age, height, **political leanings**, historical decisions, friends)

Labels: (yes or no)

- Alice liked Finding Nemo, so she'll like Zootopia
- Bob and Alice are both Republican. Bob likes this tweet $\rightarrow$ so will Alice

Mind-reading task

Features: (age, **height**, political leanings, historical decisions, friends)

Labels: (yes or no)

- Alice liked Finding Nemo, so she'll like Zootopia
- Bob and Alice are both Republican. Bob likes this tweet $\rightarrow$ so will Alice
- Claire and Dennis are both tall. Claire likes to ski $\rightarrow$ so will Dennis.

Mind-reading task

Features: (age, height, political leanings, historical decisions, friends)

Labels: (yes or no)

- Alice liked Finding Nemo, so she'll like Zootopia
- Bob and Alice are both Republican. Bob likes this tweet $\rightarrow$ so will Alice
- Claire and Dennis are both tall. Claire likes to ski $\rightarrow$ so will Dennis.

First approach: linear model

$$\underset{y}{\text{label}} = \mathbf{sign}\left(\sum_{k=1}^{d} \overset{\text{theta}}{\text{weight}_k} \times \overset{x}{\text{feature}_k}\right)$$

Training: learn weights so that prediction on training set is about right

Logistic regression model

Logistic "regression" model

- Not regression (wrongly named) but classification

june july aug.
[1, 1, 0]

- Features $x[1] =$ age, $x[2] =$ height, $x[3] =$ past viewing behavior...
- We write $x \in \mathbb{R}^d$ to mean x is a vector with d values
- Importance weights $\theta \in \mathbb{R}^d$

$$\theta[i] = \begin{cases} \text{large, positive} & = \text{feature correlates with label} \\ \text{small, around 0} & = \text{feature probably not that important} \\ \text{large, negative} & = \text{feature inversely correlates with label} \end{cases}$$

Logit model

$\theta \cdot x$

$$\text{Log likelihood} = \log\left(\frac{\Pr(y=1|x,\theta)}{\Pr(y=0|x,\theta)}\right) = \underbrace{\theta^T x := \sum_{k=1}^{d} \theta[k]x[k]}_{\text{notation for inner product}}$$

example of correlated 1-D data + labels
x1 = .1, x2 = 1., x3 = 2., x4 = -5.
y1 = 1, y2 = 1, y3 = 1, y4 = 0

Logistic regression model

- Logit model

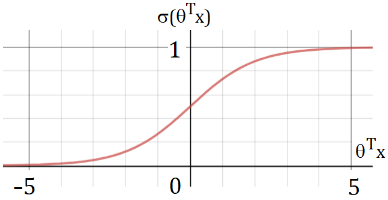
$$\text{Log likelihood} := \log\left(\frac{\Pr(y=1|x,\theta)}{\Pr(y=0|x,\theta)}\right) = x^T\theta$$

- Rearrange, normalize $\Pr(y=1) + \Pr(y=0) = 1$

$$\Pr(y=1|x,\theta) = \frac{e^{\theta^T x}}{1+e^{\theta^T x}} =: \sigma(\theta^T x)$$

- $\sigma(\theta^T x)$ is the sigmoidal function, models "soft probability"

[cat, dog, camel, bird]
1, 0, 0 0
0,1, 0 0

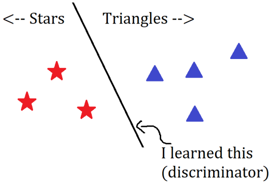


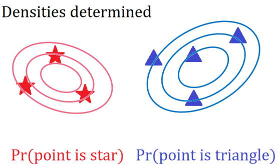
Discriminative vs generative

Logistic regression is a discriminative approach

$$\Pr(y = 1 | x, \theta) = \sigma(\theta^T x), \qquad \Pr(x | y, \theta) = \text{??? (don't care)}$$

Discriminative vs generative





Discriminative vs generative

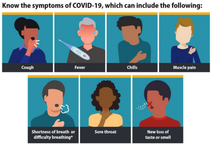
Logistic regression is a discriminative approach

$$\Pr(y = 1|x, \theta) = \sigma(\theta^T x), \qquad \Pr(x|y, \theta) = \text{??? (don't care)}$$

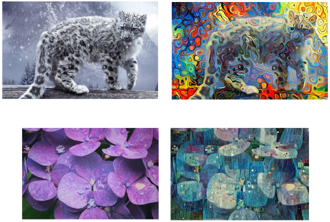
Discriminative vs generative

...roses are red,

violets are blue...



Gauge by symptoms severity



"Deep dream"
@dreamscopeapp on Twitter

Zhang et al. StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks

Discriminative vs generative

Logistic regression is a discriminative approach

$$\Pr(y = 1|x, \theta) = \sigma(\theta^T x), \qquad \Pr(x|y, \theta) = ???\ \text{(don't care)}$$

Discriminative	generative
-- if that's all the task requires (classification only) --may have less parameters ** don't need as much data	--can be used to create more training data ** depends on distribution being good -- more detailed/complete feature representation -- easy to screw up by outliers

Logistic regression

- Training samples: $\mathcal{X} = \{x_1, ..., x_m\} \subset \mathbb{R}^d$ features, $\mathcal{Y} = \{y_1, ..., y_m\}$ labels
- yes : $y_k = 1$, no : $y_k = 0$
- We assume the data is independently, identically distributed (i.i.d.)
- Logit model: $\Pr(y = 1|x) = \sigma(\theta^T x)$
- Goal: find maximum likelihood estimator

likelihood = Pr(Thing you see | underlying factor)

$$\theta_{\mathbf{MLE}} = \operatorname*{argmax}_{\theta} \cancel{\Pr(\mathcal{X}, \mathcal{Y}|\theta)}$$

Pr(Y | X, theta)

- Predictor for new sample x

$$\tilde{y}_{\text{pred}} = \mathbf{sign}(x^T \theta_{\mathbf{MLE}})$$

tilde y = 2*y-1
y = tilde y /2 + .5

{-1,1}

Derive logistic regression objective function

Pr(y = 1 | x, theta) = sigma(x.T theta)
Pr(y = 0 | x, theta) = 1 - s(x.T theta)

Likelihood of label

$$\Pr(\mathcal{Y}|\mathcal{X},\theta) \overset{\text{i.i.d.}}{=} \prod_{i=1}^{m} \Pr(y_i|x_i,\theta)$$

yi \in {0,1}

if yi = 0   if yi = 1

$$\overset{\text{Logit model}}{=} \prod_{i=1}^{m} \cancel{\sigma(\theta^T x_i)}^{y_i} (1 - \cancel{\sigma(\theta^T x_i)})^{1-y_i}$$

Maximum log likelihood

$$\max_\theta \ \log(\Pr(\mathcal{Y}|\mathcal{X},\theta)) \iff \max_\theta \sum_{i=1}^{m} \underbrace{y_i \log \sigma(\theta^T x_i)}_{\text{nonzero if } y_i=1}$$

$$+ \underbrace{(1-y_i)\log(1-\sigma(\theta^T x_i))}_{\text{nonzero if } y_i=0}, \quad y_i \in \{0,1\}$$

sigma(s) = e^s/(1 + e^s)

$$\overset{\substack{\sigma(-s)=\\ 1-\sigma(s)}}{\iff} \max_\theta \sum_{i=1}^{m} \log \sigma(\tilde{y}_i x_i^T \theta), \quad \tilde{y}_i = 2y_i - 1 \in \{-1,1\}$$

convex, 0 <= a <= 1
f (a * x + (1-a) * y) <= a * f(x) + b * f(y)
f is concave if -f is convex

x_i^T theta + 1*theta[0]   x_i = [1, ...]

Maximum likelihood estimator

y sigma(x.T theta) = sigma (y x.T theta)

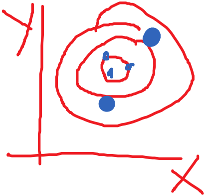
rewards

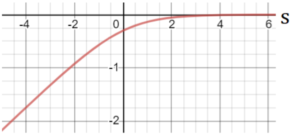
d
yi * sum xi[k] * theta[k]
k=1

$$\theta_{\mathbf{MLE}} = \underset{\theta}{\operatorname{argmax}} \left(\mathcal{L}(\theta; \mathcal{X}, \mathcal{Y}) := \sum_{i=1}^{m} \log \sigma(y_i x_i^T \theta) \right), \quad y_i \in \{-1, 1\}$$

really good margin

margin





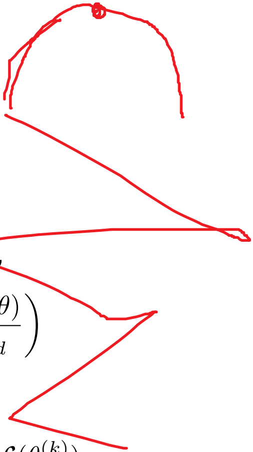
log σ(s)

- function $\mathcal{L}(\theta; \mathcal{X}, \mathcal{Y})$ is concave in θ
- $\theta_{\mathbf{MLE}}$ are the stationary points of $\mathcal{L}$

$$\theta = \theta_{\mathbf{MLE}} \iff \nabla_\theta \mathcal{L}(\theta; \mathcal{X}, \mathcal{Y}) = 0$$

if and only if

Training



$$\theta = \theta_{\mathbf{MLE}} \iff \nabla_\theta \mathcal{L}(\theta; \mathcal{X}, \mathcal{Y}) = 0$$

Gradient ascent

- For a function $\mathcal{L} : \mathbb{R}^d \to \mathbb{R}$, the gradient $\nabla \mathcal{L}(\theta) \in \mathbb{R}^d$,

$$\nabla \mathcal{L}(\theta) = \left(\frac{\partial \mathcal{L}(\theta)}{\partial \theta_1}, \frac{\partial \mathcal{L}(\theta)}{\partial \theta_2}, \cdots, \frac{\partial \mathcal{L}(\theta)}{\partial \theta_d} \right)$$

- Keep climbing up!

L(t) = 1 + t1 + t2^2, g L(t) = (1,2 t2)

$$\theta^{(0)} = \text{anywhere}, \quad \theta^{(k+1)} = \theta^{(k)} + \alpha \nabla \mathcal{L}(\theta^{(k)})$$

|| grad L (theta K)||_2 ---> 0

for a suitably small step size $\alpha > 0$

- Stop when $\nabla \mathcal{L}(\theta^{(k)})$ is close enough to 0

test misclass rate (k) - test misclass rate(k-1) \approx 0

Extensions: acceleration, using higher order derivatives, parallelization, stochastic gradients...

Logistic regression summary

given training data

- **Training**: Using data x_i, labels $y_i \in \{-1, 1\}$, find

$$\theta = \underset{\theta}{\arg\overset{\text{max}}{\cancel{\min}}}\ \cancel{-} \sum_{i=1}^{m} \log \sigma(y_i x_i^T \theta)$$

- **Prediction**: New data sample x

$$y = \begin{cases} \mathbf{sign}(\theta^T x) & \text{(hard decision)} \\ \sigma(\theta^T x) & \text{(soft decision)} \end{cases}$$

Generalization

- Predictor $\theta_{\mathbf{MLE}}$ depends cruicially on training set $\mathcal{X}, \mathcal{Y}$
- But what if training sample is not that representative?
  - Not enough data (and coverage for rate events)
  - Presence of damaging outliers
  - Data corrupted, anonymized, or tampered
- Given loss function $\mathcal{L}(\theta; x, y)$ promoting high $\Pr(y|\theta, x)$, finite training set $\mathcal{T} = \{(x_1, y_1), ...\}$

$$R^* = \underbrace{\mathbb{E}_{x,y}[\mathcal{L}(\theta; x, y)]}_{\text{Expected risk}}, \qquad R_{\mathcal{T}} = \underbrace{\frac{1}{|\mathcal{T}|} \sum_{(x,y)\in\mathcal{T}} \mathcal{L}(\theta; x_i, y_i)}_{\text{Empirical risk, training loss}}$$

$$\underbrace{R^*}_{\text{want}} = \underbrace{\overset{\text{training loss}}{R_{\mathcal{T}}}}_{\text{get}} + \underbrace{(R^* - R_{\mathcal{T}})}_{\text{Generalization loss}}$$

- Solutions?: regularization, MAP estimator, ensemble learning, more data ...

What else? We will cover

- Further analysis on logistic regression
    - multiclass classification, margin maximization method, ...
- Other classification methods
    - thresholded linear regression, support vector machines, decision trees ...
- More details on computation of training
    - how to pick step size, stochastic methods, nonconvex models ...
- Generative approaches
    - GMMs, HMMs, expectation maximization ...
- ...

Important but we will sweep under the rug

- Data balancing (rare diseases, natural disasters, car accidents)
- Data preprocessing/cleaning
- Data anonymizing/privacy
- Cost of making the wrong decision (ethical, computational)
- Other competing metrics (cost vs quality of service, retention vs addiction)