ISTA 421/521 Introduction to Machine Learning

Lecture 15: Logistic Regression

Clay Morrison

clayton@sista.arizona.edu

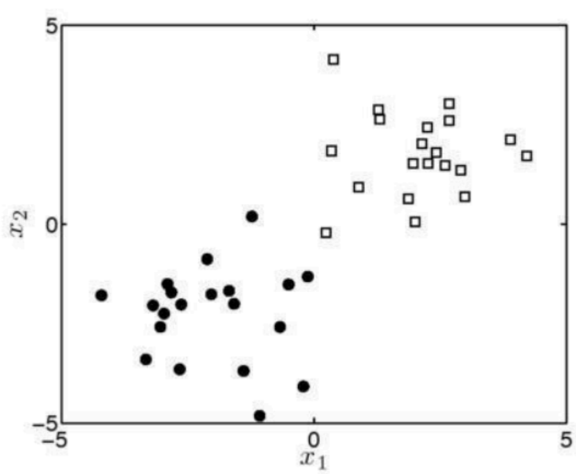
Gould-Simpson 819

Phone 621-6609

9 October 2014

Binary Classification!

- A very common type of problem
- *Many* different approaches; we'll start with a probabilistic method: logistic regression



two attributes (x_1 and x_2)

binary target, $t = \{0, 1\}$

t = 0 are dark circles
t = 1 are white squares

The Likelihood

- Assume the elements of **t** are independent, conditioned on **w**

$$p(\mathbf{t}|\mathbf{X},\mathbf{w}) = \prod_{n=1}^{N} p(t_n|\mathbf{x}_n,\mathbf{w})$$

- Previously, **t** was Gaussian distributed b/c the target was real-valued. Now the target is a binary class label (0 or 1), so likelihood is a different RV:

$$p(\mathbf{t}|\mathbf{X},\mathbf{w}) = \prod_{n=1}^{N} P(T_n = t_n|\mathbf{x}_n,\mathbf{w})$$

a binary random variable

The Likelihood

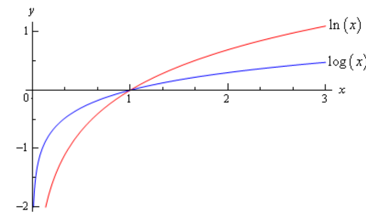
$$p(\mathbf{t}|\mathbf{X},\mathbf{w}) = \prod_{n=1}^{N} P(T_n = t_n|\mathbf{x}_n,\mathbf{w})$$

- Want likelihood to…
  - … be high if model assigns high probabilities for class 1 when we observe class 1, and high probabilities for class 0 when we observe class 0.
  - … have a maximum value of 1 where all of the training points are predicted perfectly.
- **Popular approach**: take simple linear function and pass the result through a second function that "squashes" its output, to ensure it produces a valid probability.

The Log-odds

- The logistic function is formally derived as a result of a linear model of the ***log-odds*** (aka the ***logit***):

$$\log\left(\frac{P(T_{\text{new}}=1|\mathbf{x}_{\text{new}},\mathbf{w})}{P(T_{\text{new}}=0|\mathbf{x}_{\text{new}},\mathbf{w})}\right) = \mathbf{w}^{\top}\mathbf{x}_{\text{new}}$$



- There are no constraints on this value: it can take any real value.

$P(T_{\text{new}} = 1|\mathbf{x}_{\text{new}},\mathbf{w}) \ll P(T_{\text{new}} = 0|\mathbf{x}_{\text{new}},\mathbf{w})$ Large *negative*

$P(T_{\text{new}} = 1|\mathbf{x}_{\text{new}},\mathbf{w}) \gg P(T_{\text{new}} = 0|\mathbf{x}_{\text{new}},\mathbf{w})$ Large positive

From the Logit to Logistic Function

Example of a **generalized linear model**: linear model passed through a transformation to model a quantity of interest.

- Now, derive $P(T_{\text{new}} = 1|\mathbf{x}_{\text{new}},\,\mathbf{w})$

Note: $P(T_{\text{new}}=0|\mathbf{x}_{\text{new}},\mathbf{w}) = 1 - P(T_{\text{new}}=1|\mathbf{x}_{\text{new}},\mathbf{w})$

$$\begin{aligned}
\log\left(\frac{P(T_{new}=1|\mathbf{x}_{new},\mathbf{w})}{P(T_{new}=0|\mathbf{x}_{new},\mathbf{w})}\right) &= \mathbf{w}^{\top}\mathbf{x}\\
\frac{P(T_{new}=1|\mathbf{x}_{new},\mathbf{w})}{P(T_{new}=0|\mathbf{x}_{new},\mathbf{w})} &= \exp(\mathbf{w}^{\top}\mathbf{x})\\
\frac{P(T_{new}=1|\mathbf{x}_{new},\mathbf{w})}{1-P(T_{new}=1|\mathbf{x}_{new},\mathbf{w})} &= \exp(\mathbf{w}^{\top}\mathbf{x})\\
P(T_{new}=1|\mathbf{x}_{new},\mathbf{w}) &= \exp(\mathbf{w}^{\top}\mathbf{x})(1-P(T_{new}=1|\mathbf{x}_{new},\mathbf{w}))\\
P(T_{new}=1|\mathbf{x}_{new},\mathbf{w}) &= \exp(\mathbf{w}^{\top}\mathbf{x})-\exp(\mathbf{w}^{\top}\mathbf{x})P(T_{new}=1|\mathbf{x}_{new},\mathbf{w})\\
P(T_{new}=1|\mathbf{x}_{new},\mathbf{w})+\exp(\mathbf{w}^{\top}\mathbf{x})P(T_{new}=1|\mathbf{x}_{new},\mathbf{w}) &= \exp(\mathbf{w}^{\top}\mathbf{x})\\
P(T_{new}=1|\mathbf{x}_{new},\mathbf{w})(1+\exp(\mathbf{w}^{\top}\mathbf{x})) &= \exp(\mathbf{w}^{\top}\mathbf{x})\\
P(T_{new}=1|\mathbf{x}_{new},\mathbf{w}) &= \frac{\exp(\mathbf{w}^{\top}\mathbf{x})}{1+\exp(\mathbf{w}^{\top}\mathbf{x})}\\
P(T_{new}=1|\mathbf{x}_{new},\mathbf{w}) &= \frac{1}{1+\exp(-\mathbf{w}^{\top}\mathbf{x})}
\end{aligned}$$

So the logistic function is really modeling the log-odds with a linear model!

The **Logistic** function (the inverse Logit)

Logistic as Likelihood

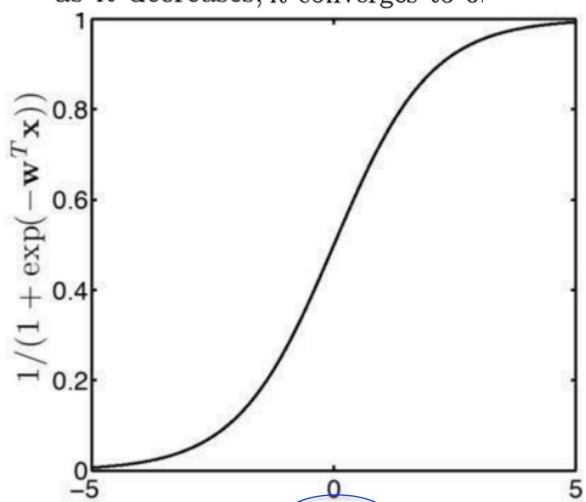
$$p(\mathbf{t}|\mathbf{X},\mathbf{w}) = \prod_{n=1}^{N} P(T_n = t_n|\mathbf{x}_n,\mathbf{w})$$

The Logistic (or Sigmoid) function

$$P(T_n = 1|\mathbf{x}_n,\mathbf{w}) = \frac{1}{1+\exp(-\mathbf{w}^\mathsf{T}\mathbf{x}_n)}$$

Linear component

As $\mathbf{w}^\mathsf{T}\mathbf{x}$ increases, the value converges to 1 as it decreases, it converges to 0.



$\mathbf{w}^T\mathbf{x}$ (Note! Not just fn of x)

When target is 0:

$$\begin{aligned} P(T_n = 0|\mathbf{x}_n,\mathbf{w}) &= 1 - P(T_n = 1|\mathbf{x}_n,\mathbf{w}) \\ &= 1 - \frac{1}{1+\exp(-\mathbf{w}^\mathsf{T}\mathbf{x}_n)} \\ &= \frac{\exp(-\mathbf{w}^\mathsf{T}\mathbf{x}_n)}{1+\exp(-\mathbf{w}^\mathsf{T}\mathbf{x}_n)}. \end{aligned}$$

Combine both into a single probability function

$$P(T_n = t_n|\mathbf{x}_n,\mathbf{w}) = P(T_n = 1|\mathbf{x}_n,\mathbf{w})^{t_n} P(T_n = 0|\mathbf{x}_n,\mathbf{w})^{1-t_n}$$

The Likelihood

$$P(T_n = t_n|\mathbf{x}_n,\mathbf{w}) = P(T_n = 1|\mathbf{x}_n,\mathbf{w})^{t_n} P(T_n = 0|\mathbf{x}_n,\mathbf{w})^{1-t_n}$$

$$\begin{aligned} p(\mathbf{t}|\mathbf{X},\mathbf{w}) &= \prod_{n=1}^{N} P(T_n = t_n|\mathbf{x}_n,\mathbf{w}) \\ &= \prod_{n=1}^{N} P(T_n = 1|\mathbf{x}_n\mathbf{w})^{t_n} P(T_n = 0|\mathbf{x}_n,\mathbf{w})^{1-t_n} \\ &= \prod_{n=1}^{N} \left(\frac{1}{1+\exp(-\mathbf{w}^\mathsf{T}\mathbf{x}_n)}\right)^{t_n} \left(\frac{\exp(-\mathbf{w}^\mathsf{T}\mathbf{x}_n)}{1+\exp(-\mathbf{w}^\mathsf{T}\mathbf{x}_n)}\right)^{1-t_n} \end{aligned}$$

Substitute in the component likelihoods to get the final likelihood function

Bayesian Logistic Regression

$$\mathbf{x}_n = \begin{bmatrix} x_{n1} \\ x_{n2} \end{bmatrix}, \mathbf{w} = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}, \mathbf{X} = \begin{bmatrix} \mathbf{x}_1^\top \\ \mathbf{x}_2^\top \\ \vdots \\ \mathbf{x}_N^\top \end{bmatrix}$$

Want to compute the posterior density over the parameters **w** of the model

$$p(\mathbf{w}|\mathbf{t},\mathbf{X}) = \frac{p(\mathbf{t}|\mathbf{X},\mathbf{w})p(\mathbf{w})}{p(\mathbf{t}|\mathbf{X})}$$

$$p(\mathbf{t}|\mathbf{X}) = \int p(\mathbf{t}|\mathbf{X},\mathbf{w})p(\mathbf{w})\ d\mathbf{w}$$

Prior: $p(\mathbf{w}) = \mathcal{N}\ (\mathbf{0},\ \sigma^2\mathbf{I})$

Likelihood:

$$p(\mathbf{t}|\mathbf{X},\mathbf{w}) = \prod_{n=1}^{N}\left(\frac{1}{1+\exp(-\mathbf{w}^\top\mathbf{x}_n)}\right)^{t_n}\left(\frac{\exp(-\mathbf{w}^\top\mathbf{x}_n)}{1+\exp(-\mathbf{w}^\top\mathbf{x}_n)}\right)^{1-t_n}$$

Prior:

$$p(\mathbf{w}|\sigma^2) = \mathcal{N}(0,\sigma^2\mathbf{I})$$

Once we have the Posterior… $P(\mathbf{w}|\mathbf{X},\mathbf{t},\sigma^2)$

… can predict the response (class) of new objects by taking the expectation with respect to this density:

$$P(t_{\text{new}} = 1|\mathbf{x}_{\text{new}},\mathbf{X},\mathbf{t}) = \mathbf{E}_{p(\mathbf{w}|\mathbf{X},\mathbf{t},\sigma^2)}\left\{\frac{1}{1+\exp(-\mathbf{w}^\mathsf{T}\mathbf{x}_{\text{new}})}\right\}$$

Problem: the posterior is not in a standard form.

The numerator is fine: just calc prior and likelihood at observations, then multiply.

It's the denominator (marginal likelihood) that is the problem: can't integrate…

$$Z^{-1} = p(\mathbf{t}|\mathbf{X},\sigma^2) = \int p(\mathbf{t}|\mathbf{X},\mathbf{w})p(\mathbf{w}|\sigma^2)\ d\mathbf{w}$$

Numerator of Bayes Theorem $$g(\mathbf{w};\mathbf{X},\mathbf{t},\sigma^2) = \overbrace{p(\mathbf{t}|\mathbf{X},\mathbf{w})}^{\text{likelihood}}\overbrace{p(\mathbf{w}|\sigma^2)}^{\text{prior}}$$

marginal Likelihood (denominator) $$Z^{-1} = p(\mathbf{t}|\mathbf{X},\sigma^2) = \int p(\mathbf{t}|\mathbf{X},\mathbf{w})p(\mathbf{w}|\sigma^2)\ d\mathbf{w}$$

The Posterior $$p(\mathbf{w}|\mathbf{X},\mathbf{t},\sigma^2) = Z^{-1}g(\mathbf{w};\mathbf{X},\mathbf{t},\sigma^2)$$

Our Options

1. Find the single value of **w** that corresponds to the highest value of the posterior. As $g(\mathbf{w};\mathbf{X},\mathbf{t},\sigma^2)$ is proportional to the posterior, a maximum of *g* will also correspond to a maximum of the posterior. Z^{-1} is not a function of **w**
2. Approximate $P(\mathbf{w}|\mathbf{X},\mathbf{t},\sigma^2)$ with some other density that we can compute analytically.
3. Sample directly from the posterior $P(\mathbf{w}|\mathbf{X},\mathbf{t},\sigma^2)$, knowing only $g(\mathbf{w};\mathbf{X},\mathbf{t},\sigma^2)$