# SPR Day 10

## Big Idea: Generalized Linear Models

| | |
|---|---|
| Linear Regression | $\mathbb{R}$ |
| Logistic Regression (also Probit Regr.) | $\{0, 1\}$ |
| Softmax / Multiclass Logistic | $\{1, 2, \ldots C\}$ |
| Poisson Regression | all non-negative integers |
| etc. | |

## In Depth: Logistic Regression for Binary Classification

- Model description
- ML estimation
  - 1st order gradient descent
  - 2nd order gradient descent (IRLS, Newton's method)

## In Depth: Bayesian Logistic Regression

Laplace Approximation

] next time

# Generalized Linear Models

Given: $N$ labeled examples of input/output pairs $(x_n, t_n)$
where $t_n \in \mathcal{Y}$ (some output space)

Goal: Pick a target distribution with PDF/PMF "$Q$" over the space $\mathcal{Y}$

Model output given input as

$$p(t_n | x_n) = Q\left(t_n \,\middle|\, g(w^T \phi(x_n))\right)$$

- $Q$: chosen distribution
- $g$: activation function (non-linear typically)
- $\phi$: feature function

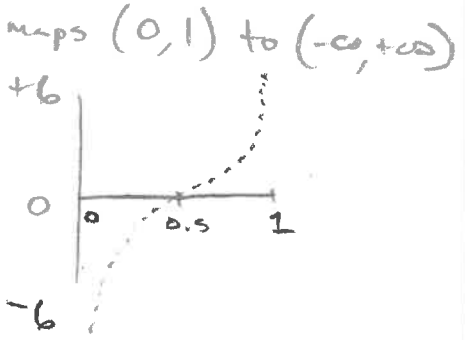
| Output Space $\mathcal{Y}$ | typical name | distribution $Q$ | activation function $g$ |
|---|---|---|---|
| real $(-\infty, +\infty)$ | linear regression | Normal | mean $= g(w^T x_n) = w^T x_n$ |
| positive real $(0, +\infty)$ | exponential regression | Exp | mean $= g(w^T x_n) = \frac{-1}{w^T x_n}$ <might have problems> |
| non-negative integer $\{0,1,2,\dots\}$ | poisson regression | Poisson | mean $= g(w^T x_n) = e^{w^T x_n}$ |
| binary $\{0,1\}$ | logistic regression | Bern | mean $= g(w^T x_n) = \sigma(w^T x_n)$ ($\sigma$: logistic sigmoid) |
| | probit regr. | Bern | $g(w^T x_n) = \Phi(w^T x_n)$ ($\Phi$: normal CDF) |
| multiple classes $\{1,2,\dots C\}$ | multi-class logistic regression | Categorical | $g(w^T x_n) = \text{softmax}(w^T x_n)$ |

# Logistic Regression for Binary Classification

Recall the logit function

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right)$$

(ln: natural log)

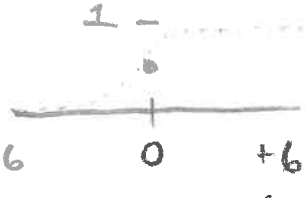



Recall the logistic sigmoid function

$$\text{sigmoid}(r) = \sigma(r) = \frac{1}{1+e^{-r}} = \frac{e^r}{e^r+1}$$

maps $(-\infty,+\infty)$ to $(0,1)$

$\sigma$: "sigma", because "S" shaped

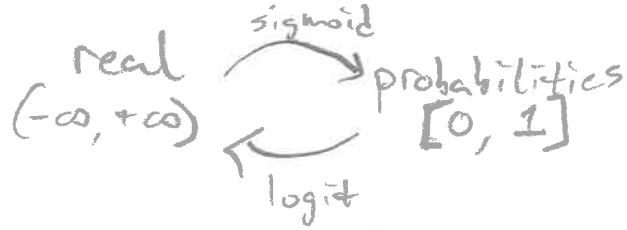



$\sigma(-6) \approx 0$

$\sigma(0) = 1/2$

$\sigma(+6) \approx 1$

note that sigmoid is inverse of logit




Model for binary outcomes:

Bernoulli w/ parameter that is monotonic transform of a linear function.

Use the sigmoid! Will produce mean parameter between 0 and 1

$$p(t_n \mid x_n) = \text{Bern}\left(t_n \mid \underbrace{\sigma(w^T\phi(x_n))}_{\text{mean}}\right)$$

$$= \sigma(w^T\phi(x_n))^{t_n}\left(1-\sigma(w^T\phi(x_n))\right)^{1-t_n}$$

$w$ is our "free" parameter: an M-dim weight vector

# Comparing ML estimates for Linear + Logistic Regression

## Linear

$$p(t_n | x_n, w) = N(t_n | w^T \phi_n, \beta^{-1})$$

ML objective:

$$\max_{w \in \mathbb{R}^M} \sum_n \log \text{NormPDF}(t_n | w^T \phi_n, \beta^{-1})$$

Solution Methods:

- Solve via analytically calculus (set deriv. to zero, solve for optima)
- Solve via 1st order gradient descent
- Solve via 2nd order gradient descent
- (others possible) like gradient-free methods (Nelder-Mead) or intelligent guess-and-check (genetic algorithms) or grid search methods

## Logistic

$$p(t_n | x_n, w) = \text{Bern}(t_n | \sigma(w^T \phi_n))$$

ML objective:

$$\max_{w \in \mathbb{R}^M} \sum_n \log \text{BernPMF}(t_n | \sigma(w^T \phi_n))$$

Solution methods:

- ~~Solve via calculus~~ Not possible + No closed form
- 1st order GD
- 2nd order GD
- (others possible)

# Gradients for Linear + Logistic Regr.

Key idea: gradient as function of $w$ looks similar across linear and logistic models

## Linear

$$\nabla_w \ell_n(w) = \nabla_w - \log \text{NormalPDF}(t_n \mid w^T \phi(x_n), \beta^{-1})$$

$$= \nabla_w \left[ +\frac{1}{2} \beta \left( t_n - \underbrace{w^T \phi(x_n)}_{\hat{y}(x_n, w)} \right)^2 \right]$$

$$= \beta \left( t_n - \hat{y}(x_n, w) \right) \cdot \nabla_w \left[ -w^T \phi(x_n) \right]$$

$$= \beta \left( \hat{y}(x_n, w) - t_n \right) \phi(x_n)$$

switch order $y$ and $t$ due to negative sign

---

## Logistic

$$\nabla_w \ell_n(w) = \nabla_w - \log \text{BernPMF}(t_n \mid \sigma(w^T x_n))$$

$$= \nabla_w \left[ -t_n \log \hat{y}(x_n, w) - (1 - t_n) \log (1 - \hat{y}(x_n, w)) \right]$$

Use $\nabla_w \log \sigma(w^T \phi_n) = \frac{1}{\sigma(w^T \phi_n)} \sigma(w^T \phi_n) \sigma(-w^T \phi_n) \phi_n$ by chain rule

$$= \left( \hat{y}(x_n, w) - t_n \right) \phi(x_n)$$

(Bishop PRML 4.91)

Useful interpretation: If predict $t_n$ perfectly, grad is zero

If overestimate $t_n$, grad positive, want to step $w$ in direction of $-$grad

# 1st and 2nd order derivatives of Maximum Lik. objective for both Linear & Logistic Regr.

$\Phi$ : N×M feature matrix

$t$ : N×1 output vector

$$\alpha = \sum_n \log p(t_n | x_n, w)$$

## Linear Regr

$$\hat{y} = \Phi w$$

predictions — actual labels

gradient $$\nabla_w \alpha = \Phi^T (\Phi w - t)$$

Hessian $$\nabla_w \nabla_w \alpha = \Phi^T \Phi$$

always constant wrt to $w$

will be positive definite ⇒ thus invertible iff $N \geq M$ and full rank

## Logistic Regression

apply sigmoid elementwise to vector $\Phi w$

gradient $$\nabla_w \alpha = \Phi^T (\sigma(\Phi w) - t)$$

Hessian $$\nabla_w \nabla_w \alpha = \Phi^T R(w) \Phi = \sum_n \hat{y}_n(w)(1 - \hat{y}_n(w)) \phi_n \phi_n^T$$

(scalars, vectors)

Let $R(w)$ be N×N diagonal matrix where entry $n,n$ is

$$[R(w)]_{nn} = \hat{y}(w, x_n)(1 - \hat{y}(w, x_n))$$

$$= \sigma(w^T \phi(x_n)) \sigma(-w^T \phi(x_n))$$

always > 0

Because $R$ has positive diagonal Log Reg. Hessian will be invertible if $\Phi^T \Phi$ is invertible and pos. definite