SPR day 9

Probabilistic Generalized Linear Models

Reading: Bishop PRML Sec 4.3

Outline:

(1) Discriminative vs. Generative

(2) Generalized Linear Models

(3) Sigmoid Function

(4) 2-class Probabilistic Logistic Regression
- ML estimation and MAP estimation
- Review possible strategies to solve $\min_{w \in \mathbb{R}^n} d(w)$

(5) Understanding 2nd order gradient methods

Discriminative vs Generative Models

Consider supervised learning task:

Given training data $\{x_n, y_n\}_{n=1}^N$ features $x_n \in \mathbb{R}^D$, labels $y_n \in \mathcal{Y}$

Learn to predict $p(y_* | x_*)$ labels-given-features

Two Approaches to probabilistic modeling

Discriminative (Sec 4.3)

Model: $p(y|x) = \prod_{n=1}^N p_w(y_n|x_n)$

only y is r.v. — x treated as fixed/known

Parameters: w generates label given feature

Prediction: directly use likelihood $p(y_*|x_*, w)$

Training: $\max_w \prod_n p(y_n|x_n, w)$

Pro:
- simpler. fewer parameters.
- directly solve supervised task

Con:
- Cannot predict if x has missing values

Generative (Sec 4.2)

Model: $p(x, y) = \prod_{n=1}^N p_\theta(x_n|y_n) p_\pi(y_n)$

both x & y as r.v.

Parameters: θ generate x given y

π generates y

Prediction via Bayes Rule: $p(y_*|x_*) = \frac{p_\theta(x_*|y_*) p_\pi(y_*)}{\sum_{y' \in \mathcal{Y}} p_\theta(x_*|y') p_\pi(y')}$

Training: $\arg\max_{\theta, \pi} \sum_n \log p_\theta(x_n|y_n) + \log p_\pi(y_n)$

Pro:
- can predict if x w/ missing values
- can sample new x from $p(x|y)$

Con:
- more complex prediction
- more parameters

Generalized Linear Models

Extend linear regression to tasks with constrained output space
(output space: real line)

Training Data: $\{x_n, t_n\}_{n=1}^N$ where $\phi(x_n) \in \mathbb{R}^M$, $t_n \in \mathcal{Y}$ (output space)

Model: $p(t_{1:N} | x_{1:N}) = \prod_{n=1}^{N} Q\left(t_n \middle| f(w^T \phi(x_n))\right)$

- Q: chosen PMF or PDF over output space $\mathcal{Y}$
- f: link function or activation function. Determines mean $\rightarrow$

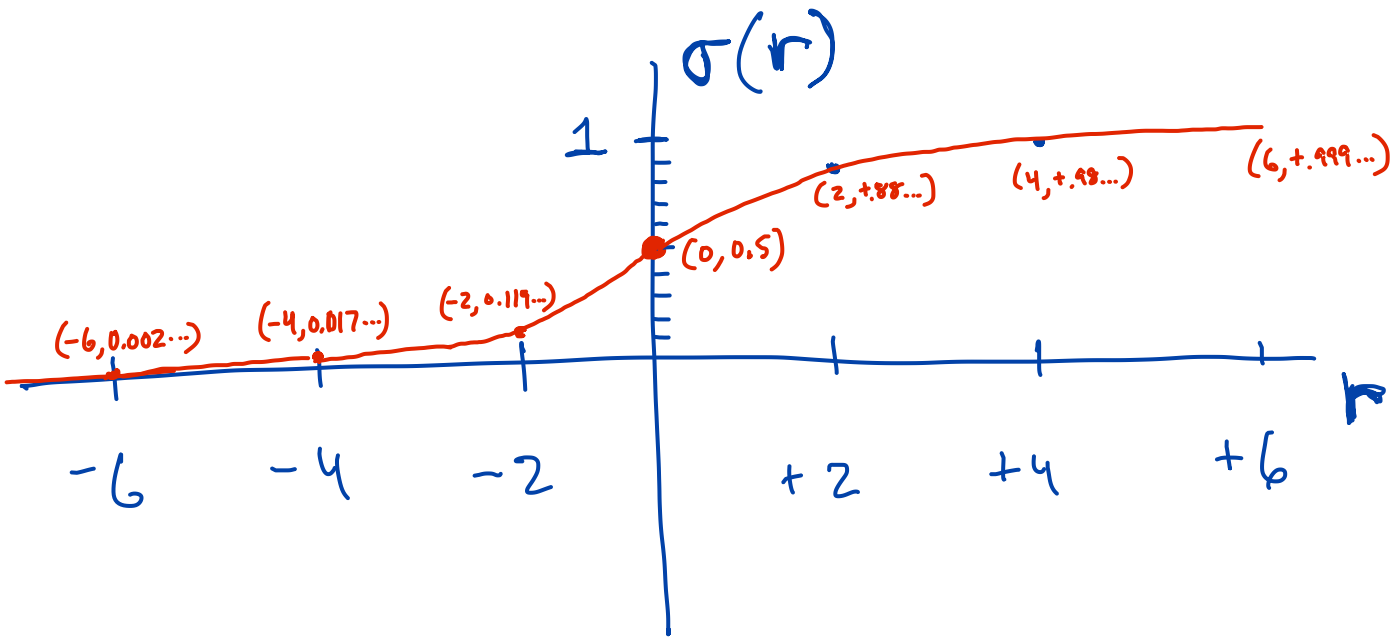
Key Idea: $\mathbb{E}[t_n] = f(w^T \phi(x_n))$

Output Space $\mathcal{Y}$	name	distribution Q	link function f
real $(-\infty, +\infty)$	linear regression	Normal	$f(w,x) = w^T \phi(x)$
integers $\{0, 1, 2, \dots\}$	Poisson regression	Poisson	$f(w,x) = e^{w^T \phi(x)}$
binary $\{0, 1\}$	Logistic regression	Bernoulli	$f(w,x) = \sigma(w^T \phi(x))$ (logistic sigmoid)
	Probit regression	Bernoulli	$f(w,x) = \Phi(w^T \phi(x))$ (Normal CDF)
C-ary classification $0\,0\,0\,\dots\,1$, $0\,0\,0\,\dots\,1\,0$, $\vdots$, $1\,0\,0\,\dots\,0\,0$	Multi-class logistic regression	Categorical	$f(w,x) = \text{Softmax}(W^T \phi(x))$ (W: $M \times C$ matrix; vector in $\mathbb{R}^C$)
C levels w/ known ranking e.g. ratings 0-5 stars; e.g. cold, warm, hot	Ordinal regression	Ordinal	look up if interested!
positive reals $(0, +\infty)$	Exponential regression	Exponential	

Sigmoid function

Let $\sigma(r)$ denote logistic sigmoid function

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



Properties:

- Maps real line to probability interval $(0,1)$
- Monotonic: $r_B > r_A$ implies $\sigma(r_B) > \sigma(r_A)$
- Invertible:

reals $(-\infty, +\infty)$ $\xrightarrow{\text{sigmoid}}$ probabilities $(0,1)$, with $\xleftarrow{\text{logit}}$

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

Probabilistic Logistic Regression

Prior: $p(w) = \mathcal{N}(w \mid m_0, S_0)$

$= \mathcal{N}(w \mid \vec{0}, \alpha^{-1} I_M)$ usually

Likelihood $p(t_n \mid w) = \text{BernPMF}(t_n \mid \sigma(w^T \phi(x_n)))$

Link function is logistic sigmoid

Monotonic transform from $(-\infty, +\infty)$ to $(0, 1)$

Point Estimation Strategies

ML estimate: $\underset{w \in \mathbb{R}^M}{\text{argmax}} \sum_{n=1}^{N} \log p(t_n \mid w)$

MAP estimate: $\underset{w \in \mathbb{R}^M}{\text{argmax}} \sum_{n=1}^{N} \log p(t_n \mid w) + \log p(w)$

Next Time: Posterior Estimation: $p(w \mid t_{1:N})$

Strategies for Optimization

$$\underset{w \in \mathbb{R}^M}{\operatorname{argmin}} \; \mathcal{L}(w)$$

convention: always minimize

	ML/MAP for Linear Regression	ML/MAP for Logistic Regr.
Analytical methods		
- Set up $\nabla_w \mathcal{L} = 0$		
- Manipulate to find $w^* = \ldots$ closed form expression	rarely works but fast + exact once implemented ✓	✗ No closed form solution exists.
Gradient methods (step size $\varepsilon_t > 0$, gradient $g \in \mathbb{R}^M$, Hessian $H \in \mathbb{R}^{M \times M}$ sym p.d.)		
- 1st order gradient descent $w_{t+1} \leftarrow w_t - \varepsilon_t g(w_t)$	Many iterations, each cheap: $O(M)$ ✓	✓
- 2nd order gradient descent $w_{t+1} \leftarrow w_t - \varepsilon_t H(w_t)^{-1} g(w_t)$	Few itrs, each expensive: $O(M^3)$ inverting matrix ✓	✓
Other methods		
- grid search	usually very inefficient in high dimensions ($M > 3$) ✓	✓
- random search	✓	✓
- Nelder-Mead	✓	✓
- ... many others possible	✓	✓

ML/MAP estimation: Gradients + Hessians

For Logistic Regression: $a(w) = -\sum_{n=1}^{N} \log \text{BernPMF}(t_n | \sigma(w^T \phi(x_n))) - \log \text{NormPDF}(w | 0, \alpha^{-1} I_M)$

	gradient $\nabla_w a$ shape: $(M, 1)$	Hessian $\nabla\nabla_w a$ shape: (M, M)
Linear Regr.	$\beta \Phi^T (\Phi w - t) + \alpha w$ (Φw: predicted, t: true label)	$\beta \Phi^T \Phi + \alpha I_M$
Logistic Regr.	$\Phi^T (\sigma(\Phi w) - t) + \alpha w$ ($\sigma(\Phi w)$: prediction (probability), t: true binary label)	$\Phi^T R(w) \Phi + \alpha I_M$

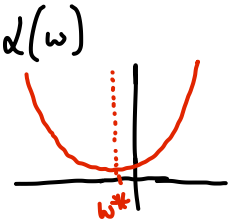
where $R(w) = \begin{bmatrix} r_1(w) & & \\ & \ddots & \\ & & r_N(w) \end{bmatrix}$ ($N \times N$ diagonal matrix)

All diagonal entries $r_n > 0$, off-diagonal entries $= 0$

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

always > 0 since $\sigma(a) > 0$ for all a

Question: Is Linear Regression MAP convex?



Unique minima w^* exists when objective is convex.

Second derivative test is positive.

In M-dims, Hessian is positive definite.

Yes. Hessian $= \alpha I_M + \Phi^T \Phi$. Sum of p.d. and psd is positive definite.

(αI_M: pos. def. if $\alpha > 0$; $\Phi^T \Phi$: p.s.d.)

Is Logistic Regression MAP convex?

Yes. R is a positive diagonal, so $\Phi^T R \Phi$ still psd, $\alpha I + \Phi^T R \Phi$ is p.d.

2nd order gradient methods

Solving an optimization problem:

$$\arg\min_{w \in \mathbb{R}^M} \mathcal{L}(w)$$

shape

Let $g(w) = \nabla_w \mathcal{L}$ $(M, 1)$

$H(w) = \nabla_w \nabla_w \mathcal{L}$ (M, M)

1st order GD

$$w^{new} \leftarrow w^{old} - \varepsilon\, g(w^{old})$$

if $M=1$ special case:

$$w^{new} \leftarrow w^{old} - \varepsilon \cdot \mathcal{L}'(w^{old})$$

2nd order GD

$$w^{new} \leftarrow w^{old} - \varepsilon\, H(w^{old})^{-1} g(w^{old})$$

$$w^{new} \leftarrow w^{old} - \varepsilon \cdot \frac{\mathcal{L}'(w^{old})}{\mathcal{L}''(w^{old})}$$

looks like Newton's method!

Linear Regression: 2nd order GD

Using formulas from previous page:

$$w^{new} = w^{old} - \varepsilon \left[\beta \Phi^T \Phi + \alpha I_M\right]^{-1} \left(\beta \Phi^T \Phi w^{old} - \beta \Phi^T t + \alpha w^{old}\right)$$

$$= w^{old} - \varepsilon \left(\beta \Phi^T \Phi + \alpha I_M\right)^{-1} \left[\left(\beta \Phi^T \Phi + \alpha I_M\right) w^{old} - \beta \Phi^T t\right]$$

$$= w^{old} - \varepsilon\, w^{old} + \varepsilon \left(\beta \Phi^T \Phi + \alpha I_M\right)^{-1} \beta \Phi^T t$$

If step size $\varepsilon = 1$, we get:

$$w^{new} = \left(\Phi^T \Phi + \frac{\alpha}{\beta} I_M\right)^{-1} \Phi^T t$$

optimal MAP estimate in one step!

Anytime loss is quadratic, 2nd order gradient update with step size $\varepsilon = 1$ will find global minima in one step.

Logistic Regression w/ 2nd order GD

Still gold standard for ML/MAP estimation, but requires many iterations (unlike linear regression ML estimation using 2nd order methods).

Update looks like this (assuming ML estimation. MAP is similar.)

$$w^{new} \leftarrow w^{old} - \varepsilon \left(\Phi^T R(w^{old}) \Phi \right)^{-1} \Phi^T \left(\underbrace{\sigma(\Phi w^{old})}_{\hat{y}} - t \right)$$

Assume $\varepsilon = 1$ and rearranging terms

$$\leftarrow \left(\Phi^T R \Phi \right)^{-1} \left(\Phi^T R \Phi w^{old} - \Phi^T \hat{y} + \Phi^T t \right)$$

$$\leftarrow \left(\Phi^T R \Phi \right)^{-1} \left(\Phi^T R z \right)$$

where $z \in \mathbb{R}^N$

$$z = \Phi w^{old} - R^{-1}(\hat{y} - t)$$

Interpret as "least squares" with per-example "weights" r, $R = \text{diag}(r)$ and outcomes z, both of which depend on previous value of w.

Thus, the name "Iteratively Re-weighted Least Squares" (IRLS)

Often used for Generalized Linear Models to do ML estimation of weights $w \in \mathbb{R}^M$

while not converged:
- update R given w
- update z given w
- update $w \leftarrow (\Phi^T R \Phi)^{-1} \Phi^T R z$