Dual of Logistic regression.

$\vec{x}_i \in \mathbb{R}^m$

Recall logistic regression:

Data points $(\vec{x}_i)$, labels $y_i = +1$ or -1.

Want: $\vec{w}^T\vec{x} + \beta = \log \frac{p(\vec{x})}{1-p(\vec{x})}$, where $p(\vec{x})$ is the probability that the data point belongs to class 1.

We showed that finding the best $\vec{w}^T, \beta$ is the same as:
(per the maximum-liklihood estimator)

$\text{Max} \prod_{i=1}^{n} P(\cancel{y}_i = y_i)$

$$\underset{\vec{w}, \beta}{\text{Maximize}} \prod_{i=1}^{n} \left(\frac{\exp(y_i(\vec{w}^T \vec{x}_i + \beta))}{1 + \exp(y_i(\vec{w}^T\vec{x} + \beta))} \right)$$

Non-convex problem.

We can take logs to turn this into a convex problem.

$$\underset{\vec{w}, \beta}{\text{Maximize}} \log \left(\prod_{i=1}^{n} \left(\frac{\exp(y_i(\vec{w}^T\vec{x}_i + \beta))}{1 + \exp(y_i(\vec{w}^T\vec{x}_i + \beta))} \right) \right)$$

$$= \underset{\vec{w}, \beta}{\text{maximize}} \sum_{i=1}^{n} \log \left(\frac{1}{1 + \exp(-y_i(\vec{w}^T\vec{x}_i + \beta))} \right)$$

$$= \underset{\vec{w}, \beta}{\text{maximize}} \sum_{i=1}^{n} -\log \left(1 + \exp(-y_i(\vec{w}^T\vec{x}_i + \beta)) \right)$$

Instead of $\max_x f(x)$, we consider $-\min_x f(x)$.

The solution x^* to the latter also solves the former.

So consider:

$$p^* = \min_{\vec{w}, \beta} \sum_{i=1}^{n} \underbrace{\log(1 + \exp(-y_i(\vec{w}^T \vec{x}_i + \beta)))}_{\log\left(\frac{1}{P(Y_i = y_i)}\right)}$$

This is a convex problem.

It is unconstrained? What is its dual?

Let $f(t) = \log(1+e^{-t})$.

Consider $\beta = 0$ for simplicity.

$$p^* = \min_{\vec{w}} \sum_{i=1}^{n} \log(1 + \exp(-y_i(\vec{w}^T \cdot \vec{x}_i)))$$

$$= \min_{\vec{w}} \sum_{i=1}^{n} \log(1 + \exp(-y_i \vec{x}_i^T \cdot \vec{w}))$$

Define $A_{n \times m} = [y_1 \vec{x}_1, y_2 \vec{x}_2, \ldots, y_n \vec{x}_n]$. and $\vec{v} = A^T \vec{w}$; $\vec{v} = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_m \end{bmatrix}$

$$= \min_{\vec{v}, \vec{w}} \sum_{i=1}^{n} \log(1 + \exp(-v_i))$$

s.t. $\vec{v} = A^T \vec{w}$

$$1 + \exp(-v_i) = \frac{1}{P(Y_i = y_i)}$$

Now we have a "constraint".

(3)

$$p^* = \min_{\vec{v}, \vec{w}} \sum_{i=1}^{n} f(v_i).$$ Convex.

$$\vec{v} = A^T \vec{w}$$

- Only constraints are linear equality + convex problem.
- If there is a feasible point, i.e. some point s.t. $\vec{v} = A^T\vec{w}$, then Slater's condition holds, i.e. $p^* = d^*$.

What is d^*?

$$L(\vec{v}, \vec{w}, \vec{\nu}) = \sum_{i=1}^{n} f(v_i) + \vec{\nu}^T (\vec{v} - A^T \vec{w})$$

$$g(\vec{\nu}) = \min_{\vec{v}, \vec{w}} L(\vec{v}, \vec{w}, \vec{\nu}) = \min_{\vec{v}} \min_{\vec{w}} L(\vec{v}, \vec{w}, \vec{\nu})$$

$$= \min_{\vec{v}} \begin{cases} -\infty & \text{if } A\vec{\nu} \neq 0. \\ \sum_{i=1}^{n} f(v_i) + \vec{\nu}^T v & \text{if } A\vec{\nu} = 0. \end{cases}$$

Consider $h(t) = f(t) + \nu_i t = \log(1+e^{-t}) + \nu_i t$.

$$\frac{dh}{dt} = \frac{1}{1+e^{-t}} (e^{-t})(-1) + \nu_i = \frac{-1}{1+e^{t}} + \nu_i$$

Setting $= 0$. $\quad \frac{1}{1+e^t} = \nu_i \Rightarrow 1+e^t = \frac{1}{\nu_i}$

$$\Rightarrow e^t = \frac{1-\nu_i}{\nu_i} \Rightarrow t = \log\left(\frac{1-\nu_i}{\nu_i}\right).$$

Note $\frac{1-\nu_i}{\nu_i} \geq 0$ only when $\nu_i \in [0,1]$.

if $\nu_i < 0$, $\lim_{t\to\infty} h(t) = -\infty$.

if $\nu_i > 1$, since $\frac{d}{dt}\log(1+e^{-t}) = \frac{-1}{1+e^t}$ has slope < -1,

$\nu_i t$ dominates as $t \to -\infty \Rightarrow \lim_{t\to -\infty} h(t) = -\infty.$

So for $v_i \in [0,1]$.

$$e^t = \frac{1-v_i}{v_i} \quad ; \quad t = \log\left(\frac{1-v_i}{v_i}\right).$$

$$h(v_i) = \log\left(1 + e^{-\log\left(\frac{1-v_i}{v_i}\right)}\right) + v_i\left(\log\left(\frac{1-v_i}{v_i}\right)\right).$$

$$= \log\left(1 + \left(\frac{v_i}{1-v_i}\right)\right) + v_i\left(\log\left(\frac{1-v_i}{v_i}\right)\right).$$

$$= \log\left(\frac{1}{1-v_i}\right) + v_i \log\left(\frac{1}{v_i}\right) + v_i \cdot \log(1-v_i).$$

$$= -v_i \log v_i - (1-v_i)\log(1-v_i)$$

= entropy of v_i

$$\therefore\ g(\vec{v}) = \begin{cases} \sum_{i=1}^{n} -v_i \log v_i - (1-v_i)\log(1-v_i) & \text{if } A\vec{v} = 0 \text{ and } \vec{v}_i \in [0,1]\ \forall i \\ -\infty & \text{otherwise.} \end{cases}$$

<u>Dual:</u> $\max_{\vec{v}} \sum_{i=1}^{n} -v_i \log v_i - (1-v_i)\log(1-v_i)$

s.t. $v_i \in [0,1]\ \forall i$

$A\vec{v} = 0$.

What is v_i? $\quad \frac{1}{P(Y_i = y_i)} = \frac{1}{1-v_i}$ $\quad$ So $v_i = P(Y_i \neq y_i)$

$$A\vec{v} = 0 \Rightarrow [\vec{x}_1 \cdots \vec{x}_n]\begin{bmatrix} y_1(1-P(Y_1=y_1)) \\ \vdots \\ y_n(1-P(Y_n=y_n)) \end{bmatrix}$$

Note $y_i(1-P(Y_i=y_i)) = 1-p(\vec{x}_i)$ if $y_i = 1$

$= p(\vec{x}_i)$ if $y_i = -1$.

Total Least Squares.

In a normal least squares problem (sometimes called OLS for ordinary least squares) we try to find $\vec{x}$ such that

$$A\vec{x} \approx \vec{b} \quad \text{and we minimize}$$

$\|\vec{e}\|_2^2 = \|A\vec{x} - \vec{b}\|_2^2$. In this formulation, we are assuming that the errors in our data are only in $\vec{b}$. But what if there are also errors in A?

$m > n$
$A \in \mathbb{R}^{m \times n}$
$\vec{b} \in \mathbb{R}^m$

i.e. we have $\underbrace{[A + \tilde{A}]}_{\text{error in } A}\vec{x} = \vec{b} + \underbrace{\tilde{\vec{b}}}_{\text{error in } b}$

$\tilde{A}$ and $\vec{\tilde{b}}$ are perturbations we do not know.

We wish to find $\vec{x}$ s.t. $[A + \tilde{A}]\vec{x} = \vec{b} + \tilde{\vec{b}}$ but all we have is A and $\vec{b}$..

How to measure the perturbations? And minimize them i.e. find A, $\vec{b}$ closest to $A + \tilde{A}$, $b + \tilde{\vec{b}}$.

$$\text{minimize} \quad \|\tilde{A} \ \tilde{\vec{b}}\|_F^2$$

$$\text{s.t.} \quad [A + \tilde{A}]\vec{x} = \vec{b} + \tilde{\vec{b}}$$

$$[A + \tilde{A}]\vec{x} = \vec{b} + \tilde{\vec{b}} \iff [A + \tilde{A} \mid \vec{b} + \tilde{\vec{b}}]\begin{bmatrix} \vec{x} \\ -1 \end{bmatrix} = 0_{m \times 1}$$

So we want $\begin{bmatrix} \vec{x} \\ -1 \end{bmatrix} \in N\Big[\underbrace{A + \tilde{A} \mid \vec{b} + \tilde{\vec{b}}}_{n+1 \text{ cols}}\Big]\Big\} m \text{ rows}$

So this matrix must be rank deficient.

Define: $[A+\tilde{A} \mid \vec{b}+\tilde{\vec{b}}] = \tilde{Z} \in \mathbb{R}^{m\times(n+1)}$

$$[A \mid \vec{b}] = Z \in \mathbb{R}^{m\times(n+1)}$$

minimize $\|Z-\tilde{Z}\|_F^2$

s.t. $\text{Rank}(\tilde{Z}) = n$

→ solution by Eckart-Young.