Fisher's linear discriminant

$$L(\mathbf{w}) = \frac{(\text{separation of projected means})^2}{\text{sum of within-class variances}} \quad (\text{want to maximize})$$

Let $\mathbf{m}_k = \frac{1}{N_k} \sum_{n \in C_k} \mathbf{x}_n$ (class means in the original space),

$m_k = \mathbf{w}^T \mathbf{m}_k$ (means in projection space),

and $y_n = \mathbf{w}^T \mathbf{x}_n$ (projections / labels of each data point).

Then the separation of projected means^2 is:

$$(m_2 - m_1)^2 = \mathbf{w}^T \underbrace{(\mathbf{m}_2 - \mathbf{m}_1)(\mathbf{m}_2 - \mathbf{m}_1)^T}_{\mathbf{S}_B} \mathbf{w}$$

and the sum of within-class variances is:

$$\sum_k \sum_{n \in C_k} (y_n - m_k)^2 = \mathbf{w}^T \underbrace{\sum_k \sum_{n \in C_k} (\mathbf{x}_n - \mathbf{m}_k)(\mathbf{x}_n - \mathbf{m}_k)^T}_{\mathbf{S}_W} \mathbf{w}$$

So the objective is:

$$L(\mathbf{w}) = \frac{\mathbf{w}^T \mathbf{S}_B \mathbf{w}}{\mathbf{w}^T \mathbf{S}_W \mathbf{w}}$$

We set the derivative to 0 to find the solution $\mathbf{w}$:

$$\frac{d}{d\mathbf{w}} L(\mathbf{w}) = \mathbf{w}^T \mathbf{S}_W \mathbf{w} \frac{d}{d\mathbf{w}} [\mathbf{w}^T \mathbf{S}_B \mathbf{w}] - \mathbf{w}^T \mathbf{S}_B \mathbf{w} \frac{d}{d\mathbf{w}} [\mathbf{w}^T \mathbf{S}_W \mathbf{w}]$$

quotient rule

$$= \mathbf{w}^T \mathbf{S}_W \mathbf{w} \; 2 \mathbf{S}_B \mathbf{w} - \mathbf{w}^T \mathbf{S}_B \mathbf{w} \; 2 \mathbf{S}_W \mathbf{w} = 0$$

$$\Rightarrow \underbrace{(\mathbf{w}^T \mathbf{S}_W \mathbf{w})}_{\text{scalar}} \mathbf{S}_B \mathbf{w} = \underbrace{(\mathbf{w}^T \mathbf{S}_B \mathbf{w})}_{\text{scalar}} \mathbf{S}_W \mathbf{w}$$

$$\Rightarrow S_W \mathbf{w} \propto S_B \mathbf{w}$$

$$\Rightarrow S_W^{-1} S_W \mathbf{w} \propto S_W^{-1} S_B \mathbf{w}$$

$$\Rightarrow \mathbf{w} \propto S_W^{-1} (\mathbf{m}_2 - \mathbf{m}_1) \underbrace{(\mathbf{m}_2 - \mathbf{m}_1)^T \mathbf{w}}_{\text{scalar}}$$

Perceptron

* If $t_n = +1$, want $\mathbf{w}^T \boldsymbol{\phi}_n > 0 \Rightarrow \overset{+1}{(t_n)} \mathbf{w}^T \boldsymbol{\phi}_n > \overset{+1}{(t_n)} 0 = 0$.

If $t_n = -1$, want $\mathbf{w}^T \boldsymbol{\phi}_n < 0 \Rightarrow \underset{-1}{(t_n)} \mathbf{w}^T \boldsymbol{\phi}_n < \underset{-1}{(t_n)} 0 \Rightarrow t_n \mathbf{w}^T \boldsymbol{\phi}_n > 0$

i.e. we want $t_n \mathbf{w}^T \boldsymbol{\phi}_n > 0$ for all samples.

* If this condition is violated for some sample, we want it to be violated by as little as possible, i.e. for misclassified samples want $t_n \mathbf{w}^T \boldsymbol{\phi}_n \leq 0$ as close to 0 as possible, i.e. want to maximize $t_n \mathbf{w}^T \boldsymbol{\phi}_n$ / minimize $- t_n \mathbf{w}^T \boldsymbol{\phi}_n$.

* $$\frac{d}{d\mathbf{w}} [- t_n \mathbf{w}^T \boldsymbol{\phi}_n] = - t_n \boldsymbol{\phi}_n$$

Bayes theorem: $$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{P(B|A)P(A)}{\sum_i P(B|A_i)P(A_i)}$$

Maximum likelihood estimation: $$\mathbf{w}^* = \underset{\mathbf{w}}{\operatorname{argmax}}\, P(\text{Data} \,|\, \mathbf{w})$$

* Example: We want to model coin tosses. The underlying model $\mathbf{w} = P(H|\mathbf{w})$.

Then $\mathbf{w}^* = \underset{\mathbf{w}}{\operatorname{argmax}}\, L(\mathbf{w})$ where $L(\mathbf{w}) = P(\{H, T, T, H, H\} | \mathbf{w}) = P(H|\mathbf{w})^{N_H} P(T|\mathbf{w})^{N_T}$.

$$\frac{d}{d\mathbf{w}} [\log L(\mathbf{w})] = 0 = \frac{d}{d\mathbf{w}} [N_H \log P(H|\mathbf{w}) + N_T \log P(T|\mathbf{w})] = \frac{d}{d\mathbf{w}} [N_H \log \mathbf{w} + N_T \log(1-\mathbf{w})]$$

$$= \frac{N_H}{\mathbf{w}} - \frac{N_T}{1-\mathbf{w}} \Rightarrow N_H - N_H \mathbf{w} - N_T \mathbf{w} = 0 \Rightarrow \mathbf{w} = \frac{N_H}{N_H + N_T} \text{ (as expected!)}$$

Logistic regression

* $P(y_i = 1 \mid \mathbf{x}_i) = \dfrac{1}{1+e^{-\mathbf{w}^T\mathbf{x}_i}} = \boxed{\sigma(\mathbf{w}^T\mathbf{x}_i)}$ where σ is the logistic sigmoid.

* Decision boundary: $P(y=1 \mid x) \overset{?}{>} P(y=0 \mid x) \Rightarrow \dfrac{P(y=1\mid x)}{P(y=0\mid x)} \overset{?}{>} 1$

$$\Rightarrow \log\left[\frac{P(y=1\mid x)}{P(y=0\mid x)}\right] \overset{?}{>} 0 \Rightarrow \log\left[\frac{P(y=1\mid x)}{1-P(y=1\mid x)}\right] \overset{?}{>} 0$$

$$\Rightarrow \log\left[\frac{\frac{1}{1+e^{-\mathbf{w}^T\mathbf{x}}}}{1-\frac{1}{1+e^{-\mathbf{w}^T\mathbf{x}}}}\right] \overset{?}{>} 0 \Rightarrow \log\left[\frac{\frac{1}{\cancel{1+e^{-\mathbf{w}^T\mathbf{x}}}}}{\frac{(\cancel{1}+e^{-\mathbf{w}^T\mathbf{x}})\cancel{-1}}{\cancel{1+e^{-\mathbf{w}^T\mathbf{x}}}}}\right] \overset{?}{>} 0 \Rightarrow$$

$$\Rightarrow \log\frac{1}{e^{-\mathbf{w}^T\mathbf{x}}} \overset{?}{>} 0 \Rightarrow -\log e^{-\mathbf{w}^T\mathbf{x}} \overset{?}{>} 0 \Rightarrow \mathbf{w}^T\mathbf{x} \overset{?}{>} 0$$

(linear classifier)

* Solution for $\mathbf{w}$: $\mathbf{w}^* = \arg\max_{\mathbf{w}} L(\mathbf{w})$ where

$$L(\mathbf{w}) = P(\text{Data} \mid \mathbf{w}) = \prod_{i=1}^N P(y_i \mid \mathbf{w}, \mathbf{x}_i) = \prod_{i=1}^N \sigma(\mathbf{w}^T\mathbf{x}_i)^{y_i}(1-\sigma(\mathbf{w}^T\mathbf{x}_i))^{(1-y_i)}$$

($y_i=1$ if pos, $y_i=0$ if neg)

$$\log L(\mathbf{w}) = \sum_{i=1}^N y_i \log\sigma(\mathbf{w}^T\mathbf{x}_i) + (1-y_i)\log(1-\sigma(\mathbf{w}^T\mathbf{x}_i))$$

$$\boxed{\frac{d L(\mathbf{w})}{d\mathbf{w}}} = 0 = \sum_{i=1}^N y_i \frac{1}{\sigma(\mathbf{w}^T\mathbf{x}_i)} \frac{d\sigma(\mathbf{w}^T\mathbf{x}_i)}{d\mathbf{w}} + (1-y_i)\frac{1}{1-\sigma(\mathbf{w}^T\mathbf{x}_i)} \frac{-d\sigma(\mathbf{w}^T\mathbf{x}_i)}{d\mathbf{w}}$$

$= \nabla_{\mathbf{w}} L(\mathbf{w})$

$$= \sum_{i=1}^N \frac{y_i - \cancel{y_i\sigma(\mathbf{w}^T\mathbf{x}_i)} - \sigma(\mathbf{w}^T\mathbf{x}_i) + \cancel{y_i\sigma(\mathbf{w}^T\mathbf{x}_i)}}{\sigma(\mathbf{w}^T\mathbf{x}_i)(1-\sigma(\mathbf{w}^T\mathbf{x}_i))} \frac{d\sigma(\mathbf{w}^T\mathbf{x}_i)}{d\mathbf{w}}$$

$$= \sum_{i=1}^N \frac{(y_i - \sigma(\mathbf{w}^T\mathbf{x}_i))}{\cancel{\sigma(\mathbf{w}^T\mathbf{x}_i)(1-\sigma(\mathbf{w}^T\mathbf{x}_i))}} \underbrace{\cancel{\sigma(\mathbf{w}^T\mathbf{x}_i)(1-\sigma(\mathbf{w}^T\mathbf{x}_i))}}_{\text{derivative of sigmoid}} \underbrace{\mathbf{x}_i}_{\text{chain rule}}$$

$$= \sum_{i=1}^N \underbrace{(y_i - \sigma(\mathbf{w}^T\mathbf{x}_i))}_{\text{prediction error} = 0 \text{ if: } y_i=1,\ P(y_i=1\mid x_i)=1 \text{ or } y_i=0,\ P(y_i=1\mid x_i)=0} \mathbf{x}_i$$

* Gradient <u>ascent</u> update: $\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} + \eta \dfrac{d L(\mathbf{w}^{(t)})}{d\mathbf{w}}$

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} + \eta \sum_{i=1}^N (y_i - \sigma(\mathbf{w}^{(t)T}\mathbf{x}_i))\,\mathbf{x}_i$$