[redacted]

scalar

Perceptron

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

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

  i.e. we want $t_n \mathbf{w}^T \phi_n > 0$ 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 $t_n \mathbf{w}^T \phi_n \leq 0$ as close to 0 as possible, i.e. want to maximize $t_n \mathbf{w}^T \phi_n$ | minimize $- t_n \mathbf{w}^T \phi_n$.

* $\dfrac{d}{d\mathbf{w}} [- t_n \mathbf{w}^T \phi_n] = - t_n \phi_n$

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

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

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

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

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

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

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

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

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

$$= \sum_{i=1}^N \underbrace{(y_i - \sigma(\mathbf{w}^T\mathbf{x}_i))}_{\text{prediction error}} \mathbf{x}_i$$

prediction error $= 0$ if: $y_i = 1,\ P(y_i=1\mid x_i) = 1$ or
$y_i = 0,\ P(y_i=1\mid x_i) = 0$

* Gradient ~~ascent~~ **ascent** update: $\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} + \eta\,\dfrac{dL(\mathbf{w}^{(t)})}{d\mathbf{w}}$

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