RVMs for classification

Lecture **17**

Consider $K=2$: $t \in \{0,1\}$.

$$y(\vec{x},\vec{w}) = \sigma(\vec{w}^T \cdot \vec{\varphi}(\vec{x}))$$

We will use a separate α_i for each w_i.

given $\vec{\alpha} = \underline{\alpha_1 \ldots \alpha_M}$, we have:

$$\underbrace{\log p(\vec{w}\,|\,\vec{t},\vec{\alpha})}_{\text{posterior}} = \underbrace{\sum_{n=1}^{N} \left[t_n \log y_n + (1-t_n)\log(1-y_n) \right]}_{\log \mathcal{L}} - \underbrace{\frac{1}{2}\vec{w}^T A \vec{w}}_{\log(\text{prior})} + \text{const}(\vec{w}) \qquad (*)$$

$$A = \begin{pmatrix} \alpha_1 & & 0 \\ & \ddots & \\ 0 & & \alpha_M \end{pmatrix}$$

Maximize $(*)$ wrt $\vec{w}$ to get $\vec{w}^*$:

$$\frac{\partial}{\partial w_j}\left[t_n \log y_n + (1-t_n)\log(1-y_n) \right] =$$

$$= \left[\frac{t_n}{y_n}\underbrace{\varphi_j(\vec{x}_n)}_{y_n(1-y_n)} - \frac{(1-t_n)}{1-y_n}\varphi_j(\vec{x}_n) \right] = \frac{t_n(1-y_n) - (1-t_n)y_n}{\underbrace{y_n(1-y_n)}_{y_n(1-y_n)}}\,\varphi_j(\vec{x}_n) =$$

↑ from $\sigma'(\vec{w}^T\vec{\varphi})$

$$= \varphi_j(\vec{x}_n)(t_n - y_n) = \varphi^T_{jn}(t_n - y_n).$$

So, $$\nabla_{\vec{w}} \log p(\vec{w}\,|\,\vec{t},\vec{\alpha})\Big|_{\vec{w}^*} = \varphi^T(\vec{t}-\vec{y}) - A\vec{w}\Big|_{\vec{w}^*} = 0,$$ so that

$$\vec{w}^* = A^{-1}\varphi^T(\vec{t}-\vec{y}).$$

Furthermore,

$$\nabla_{\vec{w}} \nabla_{\vec{w}} \log p(\vec{w}|\vec{t},\vec{\alpha}) = -\Phi^T \underbrace{B}_{N\times N} \Phi - A$$

$$B = \begin{pmatrix} y_1(1-y_1) & & 0 \\ & \ddots & \\ 0 & & y_N(1-y_N) \end{pmatrix}$$

Thus $\quad p(\vec{w}|\vec{t},\vec{\alpha}) \simeq p(\vec{w}^*|\vec{t},\vec{\alpha})\, e^{-\frac{1}{2}(\vec{w}-\vec{w}^*)\Sigma^{-1}(\vec{w}-\vec{w}^*)}$, such that

Laplace appr'n for posterior

$\Sigma^{-1} = \Phi^T B \Phi + A$, evaluated at $\vec{w}^*$

$$\underbrace{p(\vec{t}|\vec{\alpha})}_{\text{evidence for } \vec{\alpha}} = \int d\vec{w}\, p(\vec{t}|\vec{w})\, p(\vec{w}|\vec{\alpha}) \simeq$$

$$\simeq p(\vec{t}|\vec{w}^*)\, p(\vec{w}^*|\vec{\alpha})\, (2\pi)^{M/2} |\Sigma|^{1/2}.$$

Now we have

$$\frac{\partial}{\partial \alpha_i} \log p(\vec{t}|\vec{\alpha}) = \frac{\partial}{\partial \alpha_i} \log p(\vec{w}^*|\vec{\alpha}) + \frac{1}{2}\frac{\partial}{\partial \alpha_i} \log|\Sigma| =$$

$$= \frac{\partial}{\partial \alpha_i}\Big[\frac{1}{2}\underbrace{\log|A|}_{\sum_{j=1}^{M}\log \alpha_j} - \frac{1}{2}\vec{w}^{*T}\underbrace{A}_{\frac{\partial A}{\partial \alpha_i} = \mathbb{1}_{(i,i)}}\vec{w}^*\Big] - \frac{1}{2}\frac{\partial}{\partial \alpha_i}\log|\underbrace{\Sigma^{-1}}_{\Phi^T B \Phi + A}| =$$

$$= \frac{1}{2\alpha_i} - \frac{1}{2}(w_i^*)^2 - \frac{1}{2}\mathrm{Tr}\Big(\Sigma \underbrace{\frac{\partial \Sigma^{-1}}{\partial \alpha_i}}_{\mathbb{1}_{(i,i)}}\Big) =$$

$$= \frac{1}{2\alpha_i} - \frac{1}{2}(w_i^*)^2 - \frac{\Sigma_{ii}}{2} = 0.$$

Then $\alpha_i (w_i^*)^2 = \underbrace{1 - \alpha_i \Sigma_{ii}}_{\equiv\, \gamma_i}$, or

$$\alpha_i^{new} = \frac{\gamma_i}{\underbrace{(w_i^*)^2}_{\text{evaluated at } \vec{\alpha}^{old},\ \vec{w}^*}} \qquad \text{update eq'n}$$

Iterate to convergence between estimating $\begin{Bmatrix} \vec{w}^* \\ \Sigma \end{Bmatrix}$ & $\vec{\alpha}$ $\Rightarrow$ obtain sparse solution, usually much sparser than SVMs

We can also analyze α_i dependence more explicitly by using $\log p(\vec{t}\,|\,\vec{\alpha}) \equiv L(\vec{\alpha})$ & splitting off the α_i-dependent terms: $L(\vec{\alpha}) = L(\vec{\alpha}_{-i}) + \lambda(\alpha_i)$ as before

---

Finally, RVMs generalize to $K > 2$ w/out difficulties:

$$\underbrace{y_k(\vec{x})}_{\text{softmax f'n}} = \frac{e^{a_k}}{\sum_{j=1}^{K} e^{a_j}} \quad , \quad a_k = \vec{w}_k^T \vec{\phi}(\vec{x})$$