Brendan Neal 1001160236|nealbre1 2018-12-04

CSC411 Homework 7

Question 1: Representer Theorem

Part 1A

Formula for reference (expanded $||w||^2 \to w^T w$)

$$J(w) = \frac{1}{N}\sum_{i=1}^{N} L\left(y^{(i)}, t^{(i)}\right) + \frac{\lambda}{2} w^T w$$

Expanding $y^{(i)}$

$$= \frac{1}{N}\sum_{i=1}^{N} L\left(g\left(w^T \psi(x)^{(i)}\right), t^{(i)}\right) + \frac{\lambda}{2} w^T w$$

Decomposing according to hint: $w = w_\Psi + w_\perp$, where w_Ψ is the projection of w onto Ψ and $w_\perp$ is orthogonal to w_Ψ

$$= \frac{1}{N}\sum_{i=1}^{N} L\left(g\left((w_\Psi + w_\perp)^T \psi(x)^{(i)}\right), t^{(i)}\right) + \frac{\lambda}{2} (w_\Psi + w_\perp)^T (w_\Psi + w_\perp)$$

$$= \frac{1}{N}\sum_{i=1}^{N} L\left(g\left(w_\Psi^T \psi(x)^{(i)} + w_\perp^T \psi(x)^{(i)}\right), t^{(i)}\right) + \frac{\lambda}{2} (w_\Psi^T w_\Psi + 2 w_\Psi^T w_\perp + w_\perp^T w_\perp)$$

Since $w_\perp$ is orthogonal to w_Ψ and every row vector in Ψ, so their dot products reduce to 0

$$= \frac{1}{N}\sum_{i=1}^{N} L\left(g\left(w_\Psi^T \psi(x)^{(i)}\right), t^{(i)}\right) + \frac{\lambda}{2} (w_\Psi^T w_\Psi + w_\perp^T w_\perp)$$

$$\geq \frac{1}{N}\sum_{i=1}^{N} L\left(g\left(w_\Psi^T \psi(x)^{(i)}\right), t^{(i)}\right) + \frac{\lambda}{2} (w_\Psi^T w_\Psi) = J(w_\Psi)$$

Thus, we have $J(w) \geq J(w_\Psi)$ where w_Ψ is the projection of w onto Ψ.

Therefore, w_Ψ is a linear combination of $\psi\left(x^{(1)}\right)^T \dots \psi\left(x^{(N)}\right)^T$, which are the rows of Ψ.

Therefore, w_Ψ must belong to the row space of Ψ .

Brendan Neal 1001160236|nealbre1 2018-12-04

Part 1B

Final Answer:

$$\alpha_{opt} = (\mathrm{KK} + N\lambda\mathrm{K})^{-1}(K^T t)$$

Proof:

Formula for reference:

$$J(w) = \frac{1}{2N}\left||t - \Psi w|\right|^2 + \frac{\lambda}{2}\left||w|\right|$$

Rephrasing using definition of squared magnitude

$$\left||a|\right|^2 = a^T a$$

$$\left||a - b\ |\right| = a^T a - 2a^T b - b^T b$$

$$\Rightarrow J(w) = \frac{1}{2N}\left(t^T t - 2t^T \Psi w + (\Psi w)^T(\Psi w)\right) + \frac{\lambda}{2} w^T w$$

$$\Rightarrow J(w) = \frac{1}{2N}(t^T t - 2t^T \Psi w + w^T \Psi^T \Psi w) + \frac{\lambda}{2} w^T w$$

Subbing in ($w = \Psi^T \alpha, w^T = \alpha^T \Psi$)

$$\Rightarrow J(\Psi^T \alpha) = \frac{1}{2N}(t^T t - 2t^T \Psi\Psi^T \alpha + \alpha^T \Psi\Psi^T \Psi\Psi^T \alpha) + \frac{\lambda}{2}\alpha^T \Psi\Psi^T \alpha$$

Simplifying using gram matrix ($K = \Psi\Psi^T$)

$$\Rightarrow J(\alpha) = \frac{1}{2N}(t^T t - 2t^T \mathrm{K}\alpha + \alpha^T \mathrm{KK}\alpha) + \frac{\lambda}{2}\alpha^T \mathrm{K}\alpha$$

Removing fraction for simplicity (won't affect finding minimizing α)

$$\Rightarrow 2N * J(\alpha) = t^T t - 2t^T \mathrm{K}\alpha + \alpha^T \mathrm{KK}\alpha + N\lambda\alpha^T \mathrm{K}\alpha$$

Rearranging addition

$$\Rightarrow 2N * J(\alpha) = \alpha^T \mathrm{KK}\alpha + N\lambda\alpha^T \mathrm{K}\alpha - 2t^T \mathrm{K}\alpha + t^T t$$

Factoring out α, α^T for first two terms:

$$\Rightarrow 2N * J(\alpha) = (\alpha^T \mathrm{KK} + N\lambda\alpha^T \mathrm{K})\alpha - 2t^T \mathrm{K}\alpha + t^T t$$

$$\Rightarrow 2N * J(\alpha) = \alpha^T(\mathrm{KK} + N\lambda\mathrm{K})\alpha - 2t^T \mathrm{K}\alpha + t^T t$$

Dividing entire equation by 2

$$\Rightarrow N * J(\alpha) = \frac{1}{2}(\alpha^T(\mathrm{KK} + N\lambda\mathrm{K})\alpha) - t^T \mathrm{K}\alpha + \frac{1}{2}(t^T t)$$

Using hint from handout:

$$argmin_\alpha\left(\frac{1}{2}(\alpha^T A\alpha) + b^T\alpha + c\right) = -A^{-1}b$$

Where:

$$A = (\mathrm{KK} + N\lambda \mathrm{K})$$

$$[b^T = -t^T K] \Rightarrow [b = -K^T t]$$

$$c = \frac{1}{2}(t^T t)$$

Then:

$$\Rightarrow argmin_\alpha \left(\frac{1}{2}(\alpha^T (\mathrm{KK} + N\lambda \mathrm{K})\alpha) - t^T K\alpha + \frac{1}{2}(t^T t) \right) = (\mathrm{KK} + N\lambda \mathrm{K})^{-1}(K^T t)$$

Question 2: Compositional Kernels

Part 2A

Final Answer

$$\psi_S(x) = \begin{bmatrix} \psi_1(x)_1 \\ \vdots \\ \psi_1(x)_A \\ \psi_2(x)_1 \\ \vdots \\ \psi_2(x)_B \end{bmatrix}$$

Where: A, B are the respective lengths of feature maps ψ_1, ψ_2

Proof

Formulae for reference

$$k_1(x, x') = \psi_1(x) \cdot \psi_1(x')$$

$$k_2(x, x') = \psi_2(x) \cdot \psi_2(x')$$

$$k_S(x, x') = k_1(x, x') + k_2(x, x')$$

Want: ψ_S such that $k_S(x, x') = \psi_S(x) \cdot \psi_S(x')$

$$\psi_S(x) \cdot \psi_S(x') = \psi_1(x) \cdot \psi_1(x') + \psi_2(x) \cdot \psi_2(x')$$

Expanding ψ_1, ψ_2 (with respective lengths A, B)

$$\psi_S(x) \cdot \psi_S(x') = \sum_{i=1}^{A} \psi_1(x)_i \times \psi_1(x')_i + \sum_{i=1}^{B} \psi_2(x)_i \times \psi_2(x')_i$$

Express the above as the dot product of the concatenation of the feature vectors

$$\psi_S(x) \cdot \psi_S(x') = \begin{bmatrix} \psi_1(x)_1 \\ \vdots \\ \psi_1(x)_A \\ \psi_2(x)_1 \\ \vdots \\ \psi_2(x)_B \end{bmatrix} \cdot \begin{bmatrix} \psi_1(x')_1 \\ \vdots \\ \psi_1(x')_A \\ \psi_2(x')_1 \\ \vdots \\ \psi_2(x')_B \end{bmatrix}$$

$$\Rightarrow \psi_S(x) = \begin{bmatrix} \psi_1(x)_1 \\ \vdots \\ \psi_1(x)_A \\ \psi_2(x)_1 \\ \vdots \\ \psi_2(x)_B \end{bmatrix}$$

Part 2B

Final Answer

$$\psi_P(x) = vec\left(\begin{bmatrix}\psi_1(x)_1 \\ \vdots \\ \psi_1(x)_A\end{bmatrix} \times \begin{bmatrix}\psi_2(x)_1 & \cdots & \psi_2(x)_B\end{bmatrix}\right)$$

$$\psi_P(x)_{(i,j)} = \left(\psi_1(x)_i \times \psi_2(x)_j\right)$$

Proof

Formulae for reference

$$k_1(x, x') = \psi_1(x) \cdot \psi_1(x')$$

$$k_2(x, x') = \psi_2(x) \cdot \psi_2(x')$$

$$k_P(x, x') = k_1(x, x') \times k_2(x, x')$$

Want: ψ_P such that $k_P(x, x') = \psi_P(x) \cdot \psi_P(x')$

$$\psi_P(x) \cdot \psi_P(x') = \psi_1(x) \cdot \psi_1(x') \times \psi_2(x) \cdot \psi_2(x')$$

Expanding ψ_1, ψ_2 (with respective lengths A, B)

$$\Rightarrow \psi_P(x) \cdot \psi_P(x') = \left(\sum_{i=1}^{A} \psi_1(x)_i \times \psi_1(x')_i\right) \times \left(\sum_{i=1}^{B} \psi_2(x)_i \times \psi_2(x')_i\right)$$

Merging summations and rearranging multiplication

$$\Rightarrow \psi_P(x) \cdot \psi_P(x') = \left(\sum_{i=1}^{A} \sum_{j=1}^{B} \psi_1(x)_i \times \psi_1(x')_i \times \psi_2(x)_j \times \psi_2(x')_j\right)$$

$$\Rightarrow \psi_P(x) \cdot \psi_P(x') = \left(\sum_{i=1}^{A} \sum_{j=1}^{B} \left(\psi_1(x)_i \times \psi_2(x)_j\right) \times \left(\psi_1(x')_i \times \psi_2(x')_j\right)\right)$$

Note that the terms associated with x are the same as the terms associated with x' (multiplication of a ψ_1 element then a ψ_2 element).

Clearly, the (i, j)-th element of the desired kernel (represented as a matrix of size $A \times B$) must be:

$$\psi_P(x)_{(i,j)} = \left(\psi_1(x)_i \times \psi_2(x)_j\right)$$

We can represent the whole kernel as a vectorization (**flattening**) of the feature mappings' matrix multiplication:

$$\psi_P(x) = vec\left(\begin{bmatrix}\psi_1(x)_1 \\ \vdots \\ \psi_1(x)_A\end{bmatrix} \times \begin{bmatrix}\psi_2(x)_1 & \cdots & \psi_2(x)_B\end{bmatrix}\right)$$