

# Introduction to Statistical Machine Learning

## Homework 4

Yota Toyama

December 1, 2016

1.

$$\mathbf{w}^* = \underset{\mathbf{w}}{\operatorname{argmin}} \left\{ \sum_{i=1}^N (y_i - \mathbf{w} \cdot \phi(x_i))^2 \right\}$$
$$\text{s.t. } \sum_{j=1}^d w_j^2 \leq \tau$$

Using Lagrange's multiplier method,

$$\min_{\lambda} \max_{\mathbf{w}} - \sum_{i=1}^N (y_i - \mathbf{w} \cdot \phi(x_i))^2 + \lambda(\tau - \sum_{j=1}^d w_j^2)$$
$$\max_{\lambda} \min_{\mathbf{w}} \sum_{i=1}^N (y_i - \mathbf{w} \cdot \phi(x_i))^2 - \lambda(\tau - \sum_{j=1}^d w_j^2)$$
$$\frac{\partial}{\partial \mathbf{w}} \left\{ - \sum_{i=1}^N (y_i - \mathbf{w} \cdot \phi(x_i))^2 + \lambda(\tau - \sum_{j=1}^d w_j^2) \right\} = 0$$
$$2 \sum_{i=1}^N \phi(x_i)(y_i - \mathbf{w} \cdot \phi(x_i)) - 2\lambda \mathbf{w} = 0$$
$$\sum_{i=1}^N \phi(x_i)(y_i - \mathbf{w} \cdot \phi(x_i)) - \lambda \mathbf{w} = 0$$

The equation above is the same as one of L2 regularized squared loss in regression.

$\therefore$  If a proper  $\tau$  which makes  $\lambda$ s of both problems same is found, the  $w^*$ s of both problems are same.

2.

$$\begin{aligned}
\hat{y} &= \operatorname{argmax}_c \log P(\mathbf{x}, y = c) \\
&= \operatorname{argmax}_c \{ \log P(\mathbf{x}|y = c) + \log P(y = c) \} \\
&= \operatorname{argmax}_c \log P(\mathbf{x}|y = c) \\
P(\mathbf{x}|y = c; \theta) &= \prod_{j=1}^d \theta_j^{x_j} (1 - \theta_j)^{1-x_j} \\
\log P(\mathbf{x}|y = c; \theta) &= \sum_{j=1}^d x_j \log \theta_j + (1 - x_j) \log(1 - \theta_j) \\
\frac{\partial \log P(X|\mathbf{y} = c; \theta)}{\partial \theta} &= 0 \\
\sum_{i=1}^N \left\{ \frac{x_{ij}}{\theta_j} - \frac{1 - x_{ij}}{1 - \theta_j} \right\} &= 0 \\
\theta_j &= \frac{1}{N} \sum_{i=1}^N x_{ij}
\end{aligned}$$

3.

$$\begin{aligned}
\gamma_{ic} &= \frac{\pi_c P(\mathbf{x}_i; \theta_c)}{\sum_{l=1}^k \pi_l P(\mathbf{x}_i; \theta_l)} \\
&= \frac{\pi_c \prod_{j=1}^d \theta_{cj}^{x_{ij}} (1 - \theta_{cj})^{1-x_{ij}}}{\sum_{l=1}^k \pi_c \prod_{j=1}^d \theta_{lj}^{x_{ij}} (1 - \theta_{lj})^{1-x_{ij}}}
\end{aligned}$$

4.

$$\begin{aligned}
L &= E_{z_{ic} \gamma_{ic}} [\log P(X, Z; \pi, \theta)] \\
&= \sum_{i=1}^N \sum_{c=1}^k \gamma_{ic} (\log \pi_c + \log P(\mathbf{x}_i; \theta_c)) \\
&= \sum_{i=1}^N \sum_{c=1}^k \gamma_{ic} (\log \pi_c + \sum_{j=1}^d \{x_{ij} \log \theta_{lj} + (1 - x_{ij}) \log(1 - \theta_{lj})\})
\end{aligned}$$

Using Langrange multiplier method with a constraint of  $\pi_c$ ,  $\sum_{c=1}^k \pi_c = 1$ ,

$$\begin{aligned}\frac{\partial L}{\partial \theta_{lj}} &= 0 \\ \sum_{i=1}^N \gamma_{ic} \left( \frac{x_{ij}}{\theta_{lj}} - \frac{1-x_{ij}}{1-\theta_{lj}} \right) &= 0 \\ \sum_{i=1}^N \gamma_{ic} (x_{ij} - \theta_{lj}) &= 0 \\ \theta_{lj} &= \frac{\sum_{i=1}^N \gamma_{ic} x_{ij}}{\sum_{i=1}^N \gamma_{ic}}\end{aligned}$$

5. Please, see a Jupyter notebook file submitted together.

## References

- [1] Christopher M. Bishop, Pattern Recognition and Machine Learning
- [2] Discussion with Tomoki Tsujimura and Bowen Shi