# 10-701/15-781, Machine Learning: Homework 2

Eric Xing, Tom Mitchell, Aarti Singh
Carnegie Mellon University
Updated on January 12, 2010

## 1 Multiclass Classification[40pt, Ni Lao]

1. **KNN [10 pt]**
2. **Conditional Gaussian Estimation [5 pt]**

**Hint:** These two identities might be useful

$$\frac{\partial x' A x}{\partial A} = x x' \tag{1}$$

$$\frac{\partial \log |A|}{\partial A^{-1}} = -A' \tag{2}$$

**Solution:** Assume there are $n$ training samples, and define $n_y = \sum_{l:y^l=y} 1$. To model $P(y)$, we have objective function

$$L_0(\pi) = \sum_l \log P(y^l; \pi) \tag{3}$$

$$= \sum_y n_y \log \pi_y \tag{4}$$

with constraint

$$\sum_y \pi_y = \mathbf{1}'\pi = 1. \tag{5}$$

where $\pi$ is an column vector of $\pi_y$. Use the Lagrangian multiplier method

$$\frac{\partial L_0}{\partial \pi} - \lambda \frac{\partial \mathbf{1}'\pi - 1}{\partial \pi} = 0 \tag{6}$$

$$\mathbf{1}'\pi = 1 \tag{7}$$

, then we have $\pi_y = n_y/n$.

To model $P(\mathrm{x}|y)$, we have objective function

$$L_y(\Sigma_y, \mu_y) = \sum_{l:y^l=y} \log P(\mathrm{x}^l | y^l; \Sigma_y, \mu_y) \tag{8}$$

$$= \sum_{l:y^l=y} \left\{ -\frac{1}{2} \log |\Sigma_y| - (\mathrm{x}^l - \mu_y)' \Sigma^{-1} (\mathrm{x}^l - \mu_y)/2 \right\} + C, \tag{9}$$

where C is a constant. Set its gradients to zeros

$$\frac{\partial L_y}{\partial \mu_y} = 2 \sum_{l:y^l=y} \left\{ \Sigma^{-1} x^l - \Sigma^{-1} \mu_y \right\} = 0 \tag{10}$$

$$\frac{\partial L_y}{\partial \Sigma^{-1}} = \frac{1}{2} \sum_{l:y^l=y} \left\{ \Sigma' - (x^l - \mu_y)(\mathrm{x}^l - \mu_y)^T \right\} = 0 \tag{11}$$

then we have

$$\mu_y = \frac{1}{n_y} \sum_{l:y^l=y} \mathrm{x}^l \tag{12}$$

$$\Sigma_y = \frac{1}{n_y} \sum_{l:y^l=y} (\mathrm{x}^l - \mu_y)(\mathrm{x}^l - \mu_y)^T \tag{13}$$

3. **Gaussian Naive Bayes Model [5 pt]**

$$\Sigma_{y,i,i} = \frac{1}{n_y} \sum_{l:y^l=y} (\mathrm{x}_i^l - \mu_{y,i})^2 \tag{14}$$

4. **Multinomial Logistic Regression [5 pt]**

   For multinomial logistic regression

$$P(y|x;\theta) = \frac{\exp(\theta_y^T \mathrm{x})}{\sum_{i=1..K} \exp(\theta_i^T \mathrm{x})} \tag{15}$$

   where $\theta_y$ is the weight vector of the $y$-th class, and $\theta$ is a concatenation of all $\theta_y$s. We assume that $\theta_K$ is a zero vector (made up of 0s), and $x$ is already augmented by a bias feature, which always has value 1.0. Define $p^l = P(y^l|x^l;\theta)$. Then we have

$$L(\theta) = \sum_l \log p^l - \lambda|\theta|_2/2 \tag{16}$$

$$= \sum_l \{\theta_{y^l}^T \mathrm{x}^l - \log \sum_{i=1..K} \exp(\theta_i^T \mathrm{x}^l)\} - \lambda\theta^T\theta/2 \tag{17}$$

$$\frac{dL(\theta)}{d\theta_y} = \sum_l (\delta(y^l = y) - p^l)\mathrm{x}^l - \lambda\theta_y \tag{18}$$

5. **Gradient Ascent [5 pt]**

6. **Overfitting and Regularization [5 pt]**

7. **[5 pt]**