Machine Learning, Mock Exam
University of Bern

06/11/2019

- **No books, notes, computers, calculators and cellular phones are allowed.**
- **This exam has 16 points in total.**
- **There are 3 questions.**

1. **[Total 5 points]** In the constrained optimization of f

$$\min_{\omega} f(\omega) \tag{1}$$
$$g_i(\omega) \leq 0 \qquad i = 1, \dots, m \tag{2}$$
$$h_j(\omega) = 0 \qquad j = 1, \dots, l \tag{3}$$

the corresponding generalized Lagrangian is

$$\mathcal{L}(\omega, \alpha, \beta) = f(\omega) + \sum_{i=1}^{m} \alpha_i g_i(\omega) + \sum_{i=1}^{l} \beta_i h_i(\omega) \tag{4}$$

where $\alpha_i > 0, \beta_i$ are Lagrange multipliers, g_i are inequality constraints, and h_i are equality constraints. There are four conditions for the Lagrange duality theory to guarantee that the primal and dual optimal solutions coincide in a convex optimization problem. One of them is the *complementary slackness condition*, $\alpha_i^* g_i(\omega^*) = 0$, $i = 1, \dots, m$.

(a) **[3 points]** List the other three conditions.

(b) **[2 points]** What are the effects of the complementary slackness condition on the optimal linear SVM classifier? Justify your answer.
Hint: The optimal linear SVM classifier can be written as

$$x^\top w^* + b^* = \sum_{i=1}^{m} \alpha_i^* y^{(i)} x^\top x^{(i)} + b^*$$

Solution
The KKT conditions are satisfied.
Such conditions are:

(a) **[1 points]** Primal feasibility $g_i(\omega^*) \leq 0$, $i = 1, \dots, m$ and $h_i(\omega^*) = 0$, $i = 1, \dots, p$;
[1 points] dual feasibility $\alpha_i^* \geq 0$, $i = 1, \dots, m$;
[1 points] Lagrangian stationarity $\nabla_w \mathcal{L}(\omega^*, \alpha^*, \beta^*) = 0$

(b) **[2 points]** From complementary condition, $\alpha_i > 0$ define the support vectors $x^{(i)}$. These are the only vectors left in the sum in the optimal classifier.

2. **[Total 5 points]** In linear regression we are given a training set with pairs $(\mathbf{x}^{(i)}, y^{(i)})$, $i = 1, \dots, m$, and we look for a vector $\theta \in \mathbf{R}^n$ such that $y^{(i)} \approx \mathbf{x}^{(i)T}\theta$.

(a) **[2 points]** What two hypothesis on the noise are usually used to simplify the maximum likelihood formulation?

(b) **[3 points]** Derive the maximum likelihood estimate of θ given a single pair (x, y) under the hypothesis from the previous question.
Hint #1: The probability density function of a Gaussian random vector $N(0, \sigma^2 Id_n)$ is given by $p(v) = \frac{1}{\sqrt{2\pi}^n \sigma^n} \exp\left(-\frac{\|v\|^2}{2\sigma^2} \right)$
Hint #2: The problem can be rewritten as $Y = X\theta + \epsilon$ where the i-th row in Y and X are y_i and x_i respectively.

Solution

(a) **[2 points]** The samples are independent.

(b) **[3 points]** We obtain θ^* by minimizing the negative log-likelihood.

$$\begin{aligned}
\mathcal{L}(\theta|X,Y) &= -logP(Y|X,\theta) = -constant + log(exp(-\frac{\|Y - X\theta\|^2}{2\sigma^2})) \\
&= -constant + \frac{\|Y - X\theta\|^2}{2\sigma^2} \\
\nabla_{theta}\mathcal{L}(\theta|X,Y) &= \frac{1}{\sigma^2}(-X^TY + X^TX\theta) \\
\nabla_{theta^*}\mathcal{L}(\theta|X,Y) &= 0 \\
&\Rightarrow -X^TY + X^TX\theta^* = 0 \\
&\Rightarrow X^TX\theta^* = X^TY \\
&\Rightarrow \theta^* = (X^TX)^{-1}X^TY
\end{aligned}$$

3. **[Total 6 points]** Show that the Poisson distribution belongs to the exponential family.

   Hint #1. If $y \sim \text{Poisson}(\lambda)$, then $p(y;\lambda) = \frac{\lambda^y e^{-\lambda}}{y!}$

   Hint #2. Distributions in the exponential family can be written in the form of $p(y;\eta) = b(y)\exp(\eta^\top T(y) - a(\eta))$ for some η, b, T and a parameters.

   Solution.

$$\begin{aligned} p(y;\lambda) =& \frac{\exp(\log(\lambda))^y \exp(-\lambda)}{y!} \textbf{[2 points]} \\ =& \frac{1}{y!}\exp(y\log(\lambda) - \lambda) \textbf{[2 points]} \end{aligned}$$

   so we have : $\eta = \log(\lambda)$, **[.5 points]**, $b(y) = \frac{1}{y!}$ **[.5 points]**, $T(y) = y$ **[.5 points]**, and $a(\eta) = \exp(\eta))$ **[.5 points]**.