ELE 535 - Machine Learning and Pattern Recognition

Fall 2018

HOMEWORK 5: Theory

Q1 Derive the derivative, and if it exists the gradient, of the following functions.

(a) For $x \in \mathbb{R}^n$, $f(x) = \sum_{j=1}^n x_j$.

(b) For $x \in \mathbb{R}^n$, $f(x) = e^{\sum_{j=1}^n x_j}$.

(c) For $x \in \mathbb{R}^n$, $f(x) = x^T A x + a^T x + b$, where $b \in \mathbb{R}$, $a \in \mathbb{R}^n$, and $A \in \mathbb{R}^{n \times n}$.

(d) For $M \in \mathbb{R}^{n \times n}$, $f(M) = \|M\|_F^2$.

(e) For $x \in \mathbb{R}^n$, $f(x) = xx^T \in \mathbb{R}^{n \times n}$.

Q2 **Matrix Inversion Lemma.** Let $M = \begin{bmatrix} A & B \\ C & D \end{bmatrix}$ with $A \in \mathbb{R}^{p \times p}$, $D \in \mathbb{R}^{q \times q}$, $B \in \mathbb{R}^{p \times q}$, $C \in \mathbb{R}^{q \times p}$.

(a) If A and D, and at least one of S_A or S_D are invertible, derive the equality (this was partially done in class):

$$(A - BD^{-1}C)^{-1} = A^{-1} + A^{-1}B(D - CA^{-1}B)^{-1}CA^{-1}. \tag{1}$$

(b) Use part (a) to show that if A and D, and at least one of $A + BDC$ or $D^{-1} + CA^{-1}B$ are invertible, then:

$$(A + BDC)^{-1} = A^{-1} - A^{-1}B(D^{-1} + CA^{-1}B)^{-1}CA^{-1}. \tag{2}$$

Q3 **On-line least squares with mini-batch updates.** You want to solve a least squares regression problem by processing the data in small batches (mini-batches), yielding a new least squares solution after each update. Assume each mini-batch contains k training examples. Group the examples in the t-th mini-batch into the columns of $X_t \in \mathbb{R}^{n \times k}$, and the corresponding targets into the rows of $y_t \in \mathbb{R}^k$. Let $P_{t-1} = \sum_{i=1}^{t-1} X_i X_i^T \in \mathbb{R}^{n \times n}$. Assume P_{t-1}^{-1} exists and is known. Similarly, let $s_{t-1} = \sum_{i=1}^{t-1} X_i y_i \in \mathbb{R}^n$. Derive the following equations for the t-th mini-batch update:

$$\hat{y}_t \stackrel{\Delta}{=} X_t^T w_{t-1}^\star \quad \text{target prediction} \tag{3}$$

$$w_t^\star = w_{t-1}^\star + P_{t-1}^{-1} X_t [I_k + X_t^T P_{t-1}^{-1} X_t]^{-1} (y_t - \hat{y}_t) \quad \text{update } w^\star \tag{4}$$

$$P_t^{-1} = P_{t-1}^{-1} - P_{t-1}^{-1} X_t [I_k + X_t^T P_{t-1}^{-1} X_t]^{-1} X_t^T P_{t-1}^{-1} \quad \text{update } P. \tag{5}$$

How do these equations change if the mini-batches are not all the same size?

Q4 **Linear regression with vector targets.** We are given training data $\{(x_i, z_i)\}_{i=1}^m$ with input examples $x_i \in \mathbb{R}^n$ and vector targets $z_i \in \mathbb{R}^d$. Place the input examples into the columns of $X \in \mathbb{R}^{n \times m}$ and the targets into the columns of $Z \in \mathbb{R}^{d \times m}$. We want to learn a linear predictor of the vector targets $z \in \mathbb{R}^d$ of test inputs $x \in \mathbb{R}^n$. To do so, first use the training data to find:

$$W^\star = \arg \min_{W \in \mathbb{R}^{n \times d}} \quad \|Y - FW\|_F^2 + \lambda \|W\|_F^2, \tag{6}$$

where we have set $Y = Z^T$ and $F = X^T$, and we require $\lambda \geq 0$ ($\lambda = 0$ removes the ridge regularizer).

(a) Show that (7) separates into d standard ridge regression problems, each solvable separately.

(b) Without using the property in (a), set the derivative of the objective function w.r.t. W equal to zero, and find an expression for the solution $W^\star$. Is the separation property evident from this expression?

Q5 **The softmax function.** This function maps $x \in \mathbb{R}^n$ to a probability mass function $s(x)$ on n outcomes. It can be written as the composition of two functions $s(x) = q(p(x))$, where $p\colon \mathbb{R}^n \to \mathbb{R}^n_+$ and $q\colon \mathbb{R}^n_+ \to \mathbb{R}^n_+$ are defined by

$$p(x) = [e^{x_i}] \qquad q(z) = z/(\mathbf{1}^T z)$$

Here $\mathbb{R}^n_+$ denotes the positive cone $\{x \in \mathbb{R}^n\colon x_i > 0\}$. The function $p(\cdot)$ maps $x \in \mathbb{R}^n$ into the positive cone $\mathbb{R}^n_+$, and for $z \in \mathbb{R}^n_+$, $q(\cdot)$ normalizes z to a probability mass function in $\mathbb{R}^n_+$.
(a) Determine the derivative of $p(x)$ at x.
(b) Determine the derivative of $q(z)$ at z.
(c) Determine the derivative of the softmax function at x.