

Homework 0

CS229T/STATS231 (Winter 2014–2015)

Please structure your writeups hierarchically: convey the overall plan before diving into details. You should justify with words why something's true (by algebra, convexity, etc.). There's no need to step through a long sequence of trivial algebraic operations. Be careful not to mix assumptions with things which are derived. Up to two additional points will be awarded for especially well-organized and elegant solutions.

Due date: Monday, Jan. 12 (at the beginning of class)

This is a diagnostic homework and will not count towards your grade. But it should give you an idea of the types of concepts and skills required for the course, and also give you an opportunity to practice some things in case you're rusty.

1. Properties of basic machine learning functions (0 points)

Consider the prediction problem of mapping some input $x \in \mathbb{R}^d$ to output y (in regression, we have $y \in \mathbb{R}$; in classification, we have $y \in \{-1, +1\}$). A linear predictor is governed by a weight vector $w \in \mathbb{R}^d$, and we typically wish to choose w to minimize the cumulative loss over a set of training examples. Two popular loss functions for classification and regression are defined (on a single example (x, y)) as follows:

- Squared loss: $\ell(w; x, y) = \frac{1}{2}(y - w \cdot x)^2$.
- Hinge loss: $\ell(w; x, y) = \max\{1 - yw \cdot x, 0\}$.

Let's study some properties of these loss functions. These will be used throughout the entire class, so it's important to obtain a good intuition for them.

(a) Show that each loss function is convex. Hint: whenever possible, use the compositional properties of convexity (i.e., sum of two convex functions is convex, etc.).

(b) Compute the subgradient of each loss function with respect to w . Recall that the subgradient of a convex function $f(w)$ at a point w , denoted $\partial f(w)$, is a set of vectors g such that $f(w') \geq f(w) + g \cdot (w' - w)$ for all w' . If f is differentiable, $\partial f(w)$ contains exactly the gradient $\nabla f(w)$.

(c) Suppose $|y| \leq 1$ and the vectors are bounded in the L_2 norm: $\|w\|_2 \leq B$ and $\|x\|_2 \leq C$. Compute an upper bound on the absolute value of each of the loss functions $|\ell(w; x, y)|$.

(d) Under the same assumptions above, compute a bound on the L_2 norm of the subgradient of the loss function: $\sup_{g \in \partial \ell(w; x, y)} \|g\|_2$.

This this class, we'll see that many of our generalization bounds rely on control of the norms of the gradients, so it's important to get a feel for these dependencies.

2. Linear algebra / convex optimization (0 points)

(a) The L_1 norm $\|\cdot\|_1$ of a vector $v \in \mathbb{R}^n$ is defined as

$$\|v\|_1 = \sum_{i=1}^n |v_i|. \quad (1)$$

The *dual norm* $\|\cdot\|_*$ of a norm $\|\cdot\|$ is defined as

$$\|v\|_* = \sup_{\|w\| \leq 1} (v \cdot w). \quad (2)$$

Compute the dual norm of the L_1 norm.

(b) The *nuclear norm* of a matrix $A \in \mathbb{R}^{n \times n}$ is defined as $\sum_{i=1}^n |\sigma_i(A)|$, where the $\sigma_1(A), \dots, \sigma_n(A)$ are the singular values of A . Show that the nuclear norm of a positive semi-definite matrix A is equal to its trace ($\text{tr}(A) = \sum_{i=1}^n A_{ii}$). (For this reason, the nuclear norm is sometimes called the trace norm.) (Hint: use the fact that $\text{tr}(ABC) = \text{tr}(BCA)$.)

(c) For any symmetric matrix $X \in \mathbb{R}^{n \times n}$, consider the function

$$g_k(X) = \sum_{i=1}^k \lambda_i(X),$$

where $\lambda_i(X)$ denotes the i -th largest eigenvalue of X , and $1 \leq k \leq n$. Show that $g_k(\cdot)$ is a convex function. Hint: write g_k as a supremum over a family of functions.

3. Probability / statistics (0 points)

(a) Suppose you flip n independent coins, each with probability heads α . You want to ensure that you get at least one heads with probability at least $1 - \delta$. How large does n have to be (as a function of α and δ)?

(b) The *variance* of a random variable X is $\text{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2]$. Suppose X_1 and X_2 are two independent random variables with variances σ_1^2 and σ_2^2 . Compute $\text{Var}(\alpha X_1 + \beta X_2)$.

(c) Let X and Y be real-valued random variables and f be some arbitrary function. Show that the following decomposition holds:

$$\mathbb{E}[(f(X) - Y)^2] = \mathbb{E}[(f(X) - \mathbb{E}[Y | X])^2] + \mathbb{E}[\text{Var}[Y | X]].$$

where

$$\text{Var}[Y | X] = \mathbb{E}[(Y - \mathbb{E}[Y | X])^2 | X]$$

is the variance of Y conditioned on X .

(d) Recall that an exponential family is a collection of probability distributions over $x \in \mathcal{X}$ (for simplicity, assume $\mathcal{X}$ is finite):

$$p(x; \theta) = \exp\{\theta \cdot \phi(x) - A(\theta)\}, \quad (3)$$

where $\theta \in \mathbb{R}^d$ is a parameter vector, $\phi(x) \in \mathbb{R}^d$ is a feature vector, and $A(\theta) = \log \sum_{x \in \mathcal{X}} \exp\{\theta \cdot \phi(x)\}$ is the log-partition function.

Compute $\nabla A(\theta)$ and $\nabla^2 A(\theta)$ (these should have nice probabilistic interpretations). Argue that $A(\theta)$ is convex.

(e) Recall that the entropy of a probability distribution p over a finite set $\mathcal{X}$ is given by:

$$H(p) = - \sum_{x \in \mathcal{X}} p(x) \log p(x). \quad (4)$$

The entropy function is non-negative and concave. Now consider the following function:

$$F(p) = \sum_{x \in \mathcal{X}} c(x)p(x) + H(p), \quad (5)$$

where p ranges over all probability distributions over $\mathcal{X}$ and $c(x)$ is an arbitrary real-valued function on $\mathcal{X}$. Compute the p that maximizes $F(p)$.

4. Feedback (0 points)

These questions are only to help us calibrate and improve the assignments, and the responses will not impact your grade.

(a) How many hours did it take you to do this assignment?

(b) On a scale of 1 to 10, how useful was this assignment (1 was not useful, 10 was very useful)?