Homework 0
CS229T/STATS231 (Fall 2018–2019)

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 awarded for especially well-organized and elegant solutions.**

Due date: 10/03/2018, 11pm

This is a diagnostic homework and will not count towards your grade (but the bonus points do count). 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. It also will allow you to see how we grade.

1. Linear algebra (*0 points*)

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

$$\|v\|_1 = \sum_{i=1}^{n} |v_i|. \tag{1}$$

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

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

Compute the dual norm of the L_1 norm. (Here $v \cdot w$ denotes the inner product between v and w: $v \cdot w \triangleq \sum_{i=1}^{n} v_i w_i$)

b **(trace is sum of singular values)** 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 $\sigma_1(A), \ldots, \sigma_n(A)$ are the singular values of A. Show that the nuclear norm of a symmetric positive semi-definite matrix A is equal to its trace ($\mathrm{tr}(A) = \sum_{i=1}^{n} A_{ii}$). (For this reason, the nuclear norm is sometimes called the trace norm.) (Hint: use the fact that $\mathrm{tr}(AB) = \mathrm{tr}(BA)$.)

c. (*3 bonus points*) **(trace is bounded by nuclear norm)** Show that the trace of a square matrix $A \in \mathbb{R}^{n \times n}$ is always less than or equal to its nuclear norm.

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

b **(subgradients of loss functions)** Compute the subgradient of each of the two loss functions with respect to w. Recall that the subgradient of a convex function $f(w)$ at a point w, denoted $\partial f(w)$, is the set of vectors g such that $f(w') \geq f(w) + g \cdot (w' - w)$ for all w'. You may use the fact that if f is differentiable, $\partial f(w)$ is equal to the singleton set of the gradient (i.e., $\{\nabla f(w)\}$).

c **(bound subgradients)** Suppose $|y| \leq 1$ and the vectors are bounded in the L_2 norm: $\|w\|_2 \leq B$ and $\|x\|_2 \leq C$. For each of the two loss functions, compute a bound on the L_2 norm of the subgradient (i.e., bound from above the quantity $\sup_{\|w\| \leq B} \sup_{g \in \partial f(w)} \|g\|_2$). The answer may depend on B and C.

In this class, many of the generalization bounds rely on control of the norms of the gradients, so it's important to get a feel for these dependencies.

3. Probability bounds (*0 points*)

a **(independent tail bound)** Suppose you have a classifier that has a probability α of making a mistake on a random example drawn from some distribution p^*. You run the classifier on n i.i.d. examples from p^*. You want to ensure that the classifier errs at least once with probability at least $1 - \delta$. How large does n have to be (as a function of α and δ) for this to happen?

b **(asymptotics)** Suppose we have two sequences of i.i.d. random variables $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$. All $2n$ random variables are jointly independent and each random variable has mean μ and variance σ^2. Define the average difference:

$$D_n = \frac{1}{n} \sum_{i=1}^{n} (X_i - Y_i).$$

Compute the following two quantities determined by some given constant $c \in \mathbb{R}$: (i) $\lim_{n\to\infty} \mathbb{P}[D_n \geq c]$ and (ii) $\lim_{n\to\infty} \mathbb{P}[D_n \geq \frac{c}{\sqrt{n}}]$.

Express your answers in terms of the cumulative density function (CDF) Φ of the standard normal distribution ($\Phi(z) = \mathbb{P}[Z \leq z]$ for $Z \sim \mathcal{N}(0, 1)$) and the given constant c.

c. (*3 bonus points*) **(Gaussian tail bound)** Suppose $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2), \ldots, X_n \sim (\mu_n, \sigma_n^2)$ are independent Gaussian random variables. Let $Z = \sum_{i=1}^{n} X_i$. Prove that

$$\forall t > 0, \Pr\left[Z - \mathbb{E}Z \geq t\right] \leq \exp\left(-\frac{t^2}{2\sum_{i=1}^{n} \sigma_i^2}\right) \tag{3}$$

4. Moments (*0 points*)

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

b **(decomposition lemma)** Let Z be a real-valued random variable and a be a real constant. Then,

$$\mathbb{E}[(Z - a)^2] = (\mathbb{E}[Z] - a)^2 + \mathrm{Var}[Z]$$

c **(moments of mixture models)**

Suppose we have the following mixture model with K mixture components; let $\pi = (\pi_1, \ldots, \pi_K)$ be the mixing proportions satisfying $\pi_i \geq 0, \forall i$ and $\|\pi\|_1 = 1$; for each component $j \in [K] \stackrel{\text{def}}{=} \{1, \ldots, K\}$, let $\mu_j \in \mathbb{R}^d$ be its mean. Define $M = (\mu_1, \ldots, \mu_K) \in \mathbb{R}^{d \times K}$ be a matrix of all the means.

Mathematically, the model is as follows:

- $h \sim \text{Multinomial}(\pi)$ (in other words, $\mathbb{P}[h = i] = \pi_i$)
- $x_1 \mid h \sim \mathcal{N}(\mu_h, I)$ ($\mathcal{N}(\mu_h, I) \triangleq$ the Gaussian distribution with mean μ_h and identity covariance.)
- $x_2 \mid h \sim \mathcal{N}(\mu_h, I)$

Compute the $d \times d$ matrix $\mathbb{E}[x_1 x_2^\top]$; express your answer in terms of π and M.

5. Exponential families (*0 points*)

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

$$p(x;\theta) = \exp\{\theta \cdot \phi(x) - A(\theta)\}, \tag{4}$$

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.

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

$$H(p) = -\sum_{x\in\mathcal{X}} p(x) \log p(x). \tag{5}$$

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

$$F(p) = H(p) + \sum_{x\in\mathcal{X}} c(x)p(x), \tag{6}$$

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 $F(p)$.