36-755, Fall 2016
Homework 1

Due Sep 14.

1. Recall that the KL divergence between two probability measures P and Q on some measurable space $(\mathcal{X}, \mathcal{B})$ with densities p and q with respect to a common dominating measure μ is

$$K(P,Q) = \begin{cases} \int_{\mathcal{X}} p(x) \log\left(\frac{p(x)}{q(x)}\right) d\mu(x) & \text{if } P << Q \\ \infty & \text{otherwise.} \end{cases}$$

   Use Jensen inequality to show that $K(P,Q) \geq 0$ with equality if and only if $P = Q$.

2. Assume that $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$ is a parametric model over the sample space $(\mathcal{X}, \mathcal{B})$, such that $P_\theta << \mu$ for all $\theta \in \Theta$, for some σ-finite dominating measure μ. Assume also that all the P_θ's have the same support and $\theta \neq \theta'$ implies that $P_\theta \neq P_{\theta'}$. Let $\mathbb{X}_n = (X_1, \ldots, X_n) \overset{id}{\sim} P_{\theta_0}$ for some $\theta_0 \in \Theta$ and write

$$L_n(\theta; \mathbb{X}_n) = \prod_i^n p_\theta(X_i),$$

   for the likelihood function at $\theta \in \Theta$, where p_θ is the density of P_θ with respect to μ

   Use the law of large numbers to show that, for any $\theta \neq \theta_0$ in Θ,

$$\lim_{n \to \infty} \mathbb{P}\left(L_n(\mathbb{X}_n; \theta_0) > L_n(\mathbb{X}_n; \theta)\right) = 1$$

   The previous result offers an asymptotic justification of why in this case the MLE is a sensible choice. *Hint: express the inequality in term of log-likelihood ratio and show that the ratio converges in probability to $K(P_{\theta_0}, P_\theta)$.*

3. (Reading exercise. **Not to be graded for correctness, but only for effort**)
   In this problem you are essentially required to reproduce a proof that can be found in the references given below. My intention is for you to read up and understand the proof rather than trying to solve this problem on your own, which would be challenging (though you are welcome to this challenge). Let $X = (X_1, \ldots, X_d) \in \mathbb{R}^d$ be a random vector with covariance matrix Σ such that $\frac{X_i}{\sqrt{\Sigma_{i,i}}}$ is sub-Gaussian with parameter ν^2, for all $i = 1, \ldots, d$. Assume we observe n i.i.d. copies of X and compute the empirical covariance matrix $\widehat{\Sigma}$. Show that, for all $i, j \in \{1, \ldots, d\}$,

$$\mathbb{P}\left(\left|\widehat{\Sigma}_{i,j} - \Sigma_{i,j}\right| > \epsilon\right) \leq C_1 e^{-\epsilon^2 n C_2},$$

   for some constants C_1 and C_2. Conclude that

$$\max_{i,j} \left|\widehat{\Sigma}_{i,j} - \Sigma_{i,j}\right| = O_P\left(\sqrt{\frac{\log d}{n}}.\right)$$

   You may want to look these references:

   - Lemma 12 in Yuan. M. (2010). High Dimensional Inverse Covariance Matrix Estimation via Linear Programming, JMLR, 11, 2261-2286.

- Lemma 1 in Ravikumar, P., Wainwright, M.J., Raskutti, G. and Yu, B. (2011). EJS, 5, 935-980.
- Lemma A.3 in Bickel, P.J. and Levina, E. (2008). Regularized estimation of large covariance matrices, teh Annals of Statistics, 36(1), 199-227.

4. (Chebyshev-Cantelli inequality) Prove the following one sided improvement of Chebyshev's inequality: for any random variable X with finite variance σ^2 and any $t > 0$,

$$\mathbb{P}\left(X - \mathbb{E}[X] \geq t\right) \leq \frac{\sigma^2}{\sigma^2 + t^2}.$$

5. From tail bounds to (first) moment bounds.

   (a) Suppose that, for all $t > 0$,
   $$\mathbb{P}\left(|X| \geq t\right) \leq c_1 e^{-c_2 n t^a},$$
   where $a \in \{1, 2\}$. Show that
   $$\mathbb{E}\left[|X|\right] \leq c_3 n^{-1/a}$$
   and express c_3 as a function of c_1 and c_2.

   (b) (From Hoeffding/Bernstein exponential inequality to high probability bounds). Suppose that, for all $t > 0$, and some positive constants a, b, c and a non-negative constant d,
   $$\mathbb{P}\left(|X| \geq t\right) \leq a\exp\left\{-\frac{nbt^2}{c + dt}\right\}.$$
   Then show that, for any $\delta \in (0, 1)$,
   $$|X| \leq \sqrt{\frac{c}{nb} \ln \frac{a}{\delta}} + \frac{d}{nb} \ln \frac{a}{\delta},$$
   with probability at least $1 - \delta$.

6. Let X be distributed like a $N_n(0, I_n)$, where I_n is the n-dimensional identity matrix. Then, $\|X\|^2 = \sum_{i=1}^n X_i^2 \sim \chi^2_n$. Show that, for any $\epsilon \in (0, 1)$

$$\mathbb{P}\left(\left|\|X\|^2 - n\right| \geq n\epsilon\right) \leq 2e^{-n\epsilon^2/8}.$$

You can use the following fact: the moment generating function of a χ^2_n is $(1-2\lambda)^{-n/2}$ for all $\lambda < 1/2$. This results says that, in high dimensions, X is concentrated around a sphere of radius $\sqrt{n}$.

7. Suppose that $X_1, \ldots, X_n$ are such that $X_i \in SG(\sigma_i^2)$, not necessarily independent. Show that $\sum_{i=1}^n X_i \in SG(\tau^2)$ and find τ. What if $X_i \in SE(\tau_i^2, \alpha_i)$ for all i?

8. (Random Projection and the Johnson-Lindenstruass Lemma).

Suppose we have a (deterministic) vector x in $\mathbb{R}^D$ and, for $\epsilon \in (0, 1/2)$ we would like to find a random mapping $f \colon \mathbb{R}^D \to \mathbb{R}^d$, where d is smaller than D, such that

$$(1 - \epsilon)\|f(x)\|^2 \leq \|x\|^2 \leq (1 + \epsilon)\|f(x)\|^2$$

with high probability. One way is to construct a $d \times D$ matrix A with iid entries from the $N(0, 1)$ distribution and then take

$$f(x) = \frac{1}{\sqrt{d}} Ax, \quad x \in \mathbb{R}^D.$$

You can think of f as being a random projection from a high-dimensional space $\mathbb{R}^D$ into the smaller space $\mathbb{R}^d$.

Show that

(a) $\|x\|^2 = \mathbb{E}\left[\|f(x)\|^2\right]$.

(b) For each $\epsilon \in (0, 1/2)$

$$\mathbb{P}\Big(\big|\|f(x)\|^2 - \|x\|^2\big| > \epsilon\|x\|^2 \Big) < 2\exp\left\{-d/4(\epsilon^2 - \epsilon^3)\right\}$$

Proceed as follows: show that the squared norm of $\frac{\sqrt{d}f(x)}{\|x\|}$ *is equal in distribution to the sum of* d *squared standard normals, and therefore has a* χ^2_d *distribution.*
In your subsequent derivation, you may use the following facts:

(a) The mfg of a χ^2_1 *at any* $\lambda < 1/2$ *is* $(1-2\lambda)^{-1/2}$ *.*

(b) For any $\epsilon \in (0, 1/2)$*, setting* $\lambda = \frac{\epsilon}{2(1+\epsilon)} < 1/2$*, we get*

$$\frac{e^{-2(1+\epsilon)\lambda}}{1-2\lambda} = (1+\epsilon)e^{-\epsilon} < e^{-1/2(\epsilon^2-\epsilon^3)}$$

ans setting $\lambda = \frac{\epsilon}{2(1-\epsilon)} < 1/2$ *we get*

$$\frac{e^{2(1-\epsilon)\lambda}}{1+2\lambda} = (1-\epsilon)e^{\epsilon} < e^{-1/2(\epsilon^2-\epsilon^3)}$$

Using the above result, show that, if we are given n deterministic vectors $(x_1, \ldots, x_n)$ in $\mathbb{R}^D$ and we compute their projections $f(x_1), \ldots, f(x_n)$ in $\mathbb{R}^d$, we are guaranteed that the all the pairwise squared distances between the projected points are distorted by at most a factor of $\epsilon \in (0, 1/2)$ with probability at least $1-\delta$ if $d \geq \frac{4(\log(1/\delta)+2\log(n))}{\epsilon^2-\epsilon^3}$. That is,

$$\|x_i - x_j\|^2(1-\epsilon) \leq \|f(x_i) - f(x_j)\|^2 \leq \|x_i - x_j\|^2(1+\epsilon), \quad \forall i \neq j,$$

with probability at least $1-\delta$.

What is striking about this result is that the dimension D of the original space does not appear anywhere in these bounds!

This is an instance of what is also known as the Johnson-Lindenstruass Lemma, which loosely speaking, states that a random projection of n points from a high-dimensional space into a d dimensional space preserves the pairwise squared distances up to a multiplicative factor of ϵ with high probability if d is of order $\frac{\log n}{\epsilon^2}$, independently of the dimension of the original space.

Notice that instead of using independent $N(0,1)$ variables to populate A, we could have used any sub-Gaussian distribution.See the requirement that $\epsilon \in (0, 1/2)$ can be weakened). See *D. Achlioptas, Database friendly random projections: Johnson-Lindenstrauss with binary coins, Journal of Computer and System Sciences 66 (2003) 671687.*