 5/15/2014

EE 127 Final

NAME: **SID:**

The exam lasts 3 hours. The maximum number of points is 40. Notes are not allowed, except for a two-sided cheat sheet of regular format.

This booklet is 10 pages total, with extra blank spaces allotted throughout, and 2 blank pages at the end, left for you to write your answers.

There are three separate problems. All the questions in this exam can be solved independently of each other.

1. (*8 points*, Topic: Projections.) We consider a line in $\mathbf{R}^n$, described as

$$\mathcal{L} := \{x_0 + tu \;:\; t \in \mathbf{R}\},$$

where x_0, u are given n-vectors. Without loss of generality we assume that $\|u\|_2 = 1$.

(a) Find a closed-form expression for the projection $z(x)$ of an arbitrary point $x \in \mathbf{R}^n$ on the line $\mathcal{L}$.

(b) Show that the squared minimum distance D^2 from $x = 0$ to the line $\mathcal{L}$ is $D^2 = x_0^T x_0 - (u^T x_0)^2$. Why is the latter expression non-negative?

(c) On what condition on a scalar $r \geq 0$ does a sphere $\mathcal{S}_r$, with center 0 and radius r, intersect the line? Make sure to express your condition clearly in mathematical form, involving x_0 and u.

(d) Assume that the sphere $\mathcal{S}_r$ does intersect the line. Express the intersection as a segment, in the form

$$\mathcal{L} \cap \mathcal{S}_r = \{x_0 + tu \;:\; t \in [t_-, t_+]\},$$

where $t_- \leq t_+$ are scalars, which you will determine.

2. (*12 points*, Topics: machine learning, SOCP.) We consider a binary classification problem where the number of data points in the negative class is far greater than that of the positive class. To avoid too large a computational burden due to the number of negatively-labelled points, we will only rely on a very simplified information: all of those points belong to a given (hyper-) sphere:

$$\mathcal{S}_- := \{\hat{x}_- + \rho u \ : \ \|u\|_2 \leq 1\}.$$

Here, $\hat{x}_- \in \mathbf{R}^n$ is the center of the hyper-sphere, and $\rho \geq 0$ a measure of its size. We assume that all the positively-labelled points, denoted $x_i^+ \in \mathbf{R}^n$, $i = 1, \ldots, m_+$, are outside the sphere.

We consider linear separation, and will parametrize a candidate separating hyperplane as

$$\mathcal{H}(w, b) = \left\{x \ : \ w^T x + b = 0\right\},$$

where $w \in \mathbf{R}^n$, $b \in \mathbf{R}$ contain the parameters of the classifier. We will impose the following requirements on $\mathcal{H}(w, b)$:

(i) All the negatively-labelled points should be in the half-space defined by $w^T x + b \leq 0$. In other words: there are no errors on the negative class.

(ii) The number of positively-labelled points that are in the same half-space is minimal. In other words: there are as few errors as possible on the positive class.

In this exercise, we develop a convex programming approach to this problem.

(a) Show that condition (i) is implied by the sufficient condition:

$$w^T \hat{x}^- + b + \rho \|w\|_2 \leq 0.$$

Hint: you may use the Cauchy-Schwartz inequality, more precisely, the fact that, for any $w \in \mathbf{R}^n$:

$$\|w\|_2 = \max_u \ u^T w \ : \ \|u\|_2 \leq 1.$$

(b) Express the number of errors made on the positively-labelled class, in terms of (w, b) and the "error" function

$$E(z) = \begin{cases} 1 & \text{if } z \leq 0 \\ 0 & \text{otherwise.} \end{cases}$$

(c) Explain why the following problem:

$$\min_{w,b} \sum_{i=1}^{m_+} \max(0, 1 - (w^T x_i^+ + b)) \ : \ w^T \hat{x}^- + b + \rho \|w\|_2 \leq 0,$$

is a sensible heuristic to address both requirements (i) and (ii). Make sure to explain how you arrive at the formulation above.

(d) Formulate the above as an SOCP. Make sure to define the variables, the objective function, and the constraints, precisely.

3. (*20 points*, Topic: SVD.) We are given two data sets encoded in matrices $X = [x_1, \ldots, x_m]$ and $Y = [y_1, \ldots, y_m]$, where $x_i, y_i \in \mathbf{R}^n$, $i = 1, \ldots, m$. We would like to analyze the correlations between these two multi-dimensional data sets. We will use the sample covariance matrices associated with X, Y:

$$\Sigma_{xx} := \frac{1}{m}\sum_{i=1}^{m}(x_i - \hat{x})(x_i - \hat{x})^T, \quad \Sigma_{yy} := \frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y})(y_i - \hat{y})^T,$$

where $\hat{x} = (1/m)(x_1 + \ldots + x_m) \in \mathbf{R}^n$ is the sample average of x, and $\hat{y}$ is defined similarly. We will also use the so-called cross-covariance matrix, which is a $n \times n$ matrix defined as

$$\Sigma_{xy} := \frac{1}{m}\sum_{i=1}^{m}(x_i - \hat{x})(y_i - \hat{y})^T.$$

Note that Σ_{xy} is not symmetric in general. We assume that Σ_{xy} is non-zero, and that Σ_{xx}, Σ_{yy} are both positive-definite.

(a) Define the *sample correlation* between two non-zero vectors $\alpha \in \mathbf{R}^m$ and $\beta \in \mathbf{R}^m$, as

$$\mathrm{Corr}(\alpha, \beta) := \frac{1}{m\sigma_\alpha\sigma_\beta}\sum_{i=1}^{m}(\alpha_i - \hat{\alpha})(\beta_i - \hat{\beta}),$$

where $\hat{\alpha}, \hat{\beta}$ are the sample averages:

$$\hat{\alpha} = \frac{1}{m}(\alpha_1 + \ldots + \alpha_m), \quad \hat{\beta} = \frac{1}{m}(\beta_1 + \ldots + \beta_m),$$

and $\sigma_\alpha^2, \sigma_\beta^2$ are the sample variances:

$$\sigma_\alpha^2 := \frac{1}{m}\sum_{i=1}^{m}(\alpha_i - \hat{\alpha})^2, \quad \sigma_\beta^2 := \frac{1}{m}\sum_{i=1}^{m}(\beta_i - \hat{\beta})^2.$$

Explain precisely why we always have $|\mathrm{Corr}(\alpha, \beta)| \leq 1$.

(b) On what condition on α, β do we have $|\mathrm{Corr}(\alpha, \beta)| = 1$?

(c) On to the multi-dimensional case: we seek linear combinations of the data points:

$$\alpha = u^T X, \quad \beta = v^T Y,$$

where the vector variables $u, v \in \mathbf{R}^n$, $u \neq 0$, $v \neq 0$, are to be chosen such that the sample correlation between the (row) vectors α, β is maximal. Express the problem as a maximization problem, of the form:

$$\max_{u\neq 0,\, v\neq 0} \frac{u^T\Sigma_{xy}v}{\sqrt{u^T\Sigma_{xx}u}\sqrt{v^T\Sigma_{yy}v}}.$$

(d) Show that the problem can be equivalently expressed as

$$\max_{u,v} \; u^T \Sigma_{xy} v \; : \; u^T \Sigma_{xx} u = 1, \;\; v^T \Sigma_{yy} v = 1. \tag{1}$$

(e) Explain how to reduce the above problem to the case when both matrices Σ_{xx}, Σ_{yy} are the identity. Make sure to describe precisely the relations between u, v and any transformed variables you are using. *Hint:* for a positive-definite matrix S, we can define the *matrix square-root* of S, denoted $S^{1/2}$, as the symmetric matrix with the same system of eigenvectors, and eigenvalues set at the square-root of the eigenvalues of S; by construction, that matrix satisfies $S^{1/2} \cdot S^{1/2} = S$.

(f) Explain how to solve problem (1) via the SVD of an appropriate matrix, which you will determine. Make sure to clearly state the optimal value of the problem, and optimal points u, v.

(g) (5-point bonus question) Consider the problem with inequality constraints instead of inequalities:

$$\phi := \max_{u,v} \; u^T \Sigma_{xy} v \; : \; u^T \Sigma_{xx} u \le 1, \;\; v^T \Sigma_{yy} v \le 1.$$

Explain why the above problem has the same optimal value and optimal points as problem (1). *Hint:* argue first that $\phi > 0$, and then prove that both inequalities are equalities at optimum, using a scaling argument.

EXTRA SPACE