CS249 – Principles of Data Mining/Data Science – Spring 2014
Midterm Exam Outline

The Midterm will have two parts:

- an in-class written exam (open book/laptop), Tuesday April 29, 4-6pm
- a take-home exam, due Saturday May 3, 11:59pm (with upload to CourseWeb).

Each of these parts will be worth 50%.
The in-class written exam will have four questions, on the following topics:

- Distributions: multidimensional gaussians, covariance and correlation matrices, Maximum Likelihood Estimation

$$\begin{aligned}
g(\boldsymbol{x}, \boldsymbol{\mu}, \Sigma) &= \frac{1}{(2\,\pi)^{p/2}}\,\frac{1}{\sqrt{\det\,\Sigma}}\,\exp\left(-\tfrac{1}{2}\,(\boldsymbol{x}-\boldsymbol{\mu})'\,\Sigma^{-1}\,(\boldsymbol{x}-\boldsymbol{\mu})\right). \\
\text{Likelihood}[\,\boldsymbol{\theta}\mid D\,] &= \text{Prob}[\,D\mid\boldsymbol{\theta}\,] = \prod_{i=1}^{n}\text{Prob}[\,d_i\mid\boldsymbol{\theta}\,] \quad \text{where} \quad D = \{\,d_1,\,\ldots,\,d_n\,\} \\
\log\,\text{Likelihood}[\,\boldsymbol{\theta}\mid D\,] &= \sum_{i=1}^{n}\log\left(\,\text{Prob}[\,d_i\mid\boldsymbol{\theta}\,]\,\right).
\end{aligned}$$

  Reading: Ricci, Fitting Distributions with R (http://cran.r-project.org/doc/contrib/Ricci-distributions-en.pdf)

- SVD, LSI, PCA, Pseudoinverses.

$$\begin{aligned}
\text{pseudoinverse} \qquad X^{-} &= V\;S^{-}\;U' \quad \text{if } X = U\,S\,V' \\
&= (X'\;X)^{-1}\;X' \quad \text{if } X'\;X \text{ is nonsingular} \\
\text{covariance matrix} \qquad \text{cov}(X) &= \frac{1}{(n-1)}\,(X-\overline{X})'\,(X-\overline{X}) \\
&= D\;\;\text{corr}(X)\;\;D.
\end{aligned}$$

  where $D = \text{diag}(\sigma_1, \ldots, \sigma_p)$ is the diagonal matrix of X's column standard deviations.

  Reading: PCA is covered in Sections 3.4.1, 3.5.1, and 14.5 in [ESL]; Sections 6.3.1 and 10.2 in [ISL]; the other topics are basic and require another source.

- Linear Regression: Least squares, Pseudoinverses.

$$\begin{aligned}
\widehat{\boldsymbol{\beta}} &= X^{-}\,\boldsymbol{y} = (X'\,X)^{-1}\,X'\,\boldsymbol{y} \\
H &= X\,(X'\,X)^{-1}\,X' \\
\widehat{\boldsymbol{y}} &= \widehat{f}(X) = X\,\widehat{\boldsymbol{\beta}} = X\,(X'\,X)^{-1}\,X'\,\boldsymbol{y} = H\,\boldsymbol{y} \\
\boldsymbol{\epsilon} &= \boldsymbol{y} - \widehat{\boldsymbol{y}} = \boldsymbol{y} - \widehat{f}(X) = \boldsymbol{y} - X\,\widehat{\boldsymbol{\beta}} \\
RSS(\boldsymbol{\beta}) &= \|\boldsymbol{\epsilon}\|^2 = \left\|\boldsymbol{y} - X\,\widehat{\boldsymbol{\beta}}\right\|^2 = (\boldsymbol{y} - X\,\widehat{\boldsymbol{\beta}})'\,(\boldsymbol{y} - X\,\widehat{\boldsymbol{\beta}})
\end{aligned}$$

  Reading: Section 3.2 in [ESL].

- Linear Classification: multidimensional gaussians, LDA, QDA.
  $\text{Prob}[\,\text{class} = k\,|\,\boldsymbol{x}\,] > \text{Prob}[\,\text{class} = \ell\,|\,\boldsymbol{x}\,]. \iff d_k > d_\ell$ where:

$$\begin{aligned}
\text{LDA rule:} \qquad d_k &= \log\,p_k - \tfrac{1}{2}\,\boldsymbol{\mu}_k{}'\,\Sigma^{-1}\boldsymbol{\mu}_k + \boldsymbol{x}'\,\Sigma^{-1}\boldsymbol{\mu}_k \\
\text{QDA rule:} \qquad d_k &= \log\,p_k - \tfrac{1}{2}\,\log\,\det(\Sigma_k) - \tfrac{1}{2}\,\boldsymbol{\mu}_k{}'\,\Sigma_k^{-1}\boldsymbol{\mu}_k + (\boldsymbol{x}-\boldsymbol{\mu}_k)'\,\Sigma_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)
\end{aligned}$$

  Reading: Section 4.4 in [ISL]; section 4.3 in [ESL].

Each of these topics centers around a few equations, notably the ones shown above. The midterm will assume you are comfortable with these and related equations in the reading.

The take-home exam will involve three or four problems using datasets covered in the HW assignments:

- Medicare payments
- Baby names
- Data-Mining-Topic Interests of CS249 students
- Some dataset from [ISL] or [ESL] (e.g., Spam).

Each problem will involve some concept related to the topics covered in the assignments.

Sample Questions

The questions will emphasize understanding of the topics, equations, and reading listed above. The format may include true/false, multiple choice, or others, but (like the equations above) will generally emphasize linear algebra. Here are some examples:

- Is true that the determinant of a covariance matrix must be positive?
- Assuming that $\boldsymbol{e}$ is an eigenvector of a matrix X, is $-\boldsymbol{e}$ an eigenvector of X?
- Show corr(X) $=$ cov(Z) where Z is the z-score matrix $Z = (X - \overline{X}) = (\,(x_{ij} - \mu_j)/\sigma_j\,)$.
- with the SVD $X = U\,S\,V'$, LSI considers the k-th degree approximation

$$X^{(k)} \;=\; U\;S^{(k)}\;V' \;=\; U^{(k)}\;S^{(k)}\;V^{(k)'}$$

  where $S^{(k)}$ is the result of setting all diagonal elements to zero after the first k entries.
- Suppose the SVD of $X \;=\; U\,S\,V'$. What are the eigenvalues of $X'\,X$?
- Given an input $\boldsymbol{x}$, suppose that

$$\text{Prob}[\text{ class} = k \mid \boldsymbol{x}\;] \;=\; \frac{g_k(\boldsymbol{x})}{\sum_\ell g_\ell(\boldsymbol{x})}.$$

  where g_k is the gaussian characterizing the k-th class. Using the definition of g, expand and simplify the expression

$$\log\,\frac{\text{Prob}[\text{ class} = k \mid \boldsymbol{x}\;]}{\text{Prob}[\text{ class} = \ell \mid \boldsymbol{x}\;]} \;=\; \log\,\frac{g_k(\boldsymbol{x})}{g_\ell(\boldsymbol{x})}.$$

- The iris dataset has 3 classes of iris. Suppose we want to create a QDA classifier. The usual presentation of QDA considers only 2 classes; how can we implement a 3-class classifier for irises using QDA?
- Let X be a unitary matrix. What is its pseudoinverse X^-?
- Consider the 928×2 Galton dataset $G = (\,\boldsymbol{x} \mid \boldsymbol{y}\,)$, containing pairs of points (x_i, y_i) where x_i is the height of the i-th parent, and y_i is the height of the i-th child. Assuming there is no intercept, give an equation for the regression line through G.
- With the Galton dataset, $\overline{\boldsymbol{x}} \approx \overline{\boldsymbol{y}} \approx 68$. However the standard deviations differ: $\sigma_x = 7.4$, while $\sigma_y = 2.2$. How could the means be equal but the standard deviations be so different?
- With the Galton dataset, the covariance matrix and its SVD are approximately

$$\begin{pmatrix} 3.2 & 2.1 \\ 2.1 & 6.3 \end{pmatrix} \;=\; \begin{pmatrix} 0.44 & 0.90 \\ 0.90 & -0.44 \end{pmatrix} \begin{pmatrix} 7.35 & \\ & 2.17 \end{pmatrix} \begin{pmatrix} 0.44 & 0.90 \\ 0.90 & -0.44 \end{pmatrix}.$$

  Is the first principal component equivalent to the regression line through the data?
- Show that for the least squares model $\boldsymbol{y} \;\sim\; \beta_0 \;+\; \boldsymbol{x}\,\beta_1$ the least squares solution is

$$\beta_1 \;=\; \frac{\text{cov}(\boldsymbol{x},\boldsymbol{y})}{\text{cov}(\boldsymbol{x},\boldsymbol{x})}, \qquad \beta_0 \;=\; \overline{\boldsymbol{y}} \;-\; \overline{\boldsymbol{x}}\,\beta_1.$$

- Are the 3 classes in the iris dataset well-described by 4D gaussians? If so, do they have the same covariance matrix?
- Sketch the points of a 2D dataset on which LDA gives better accuracy than QDA.