Gaussian and Linear Classifiers

goo.gl/txOkp2

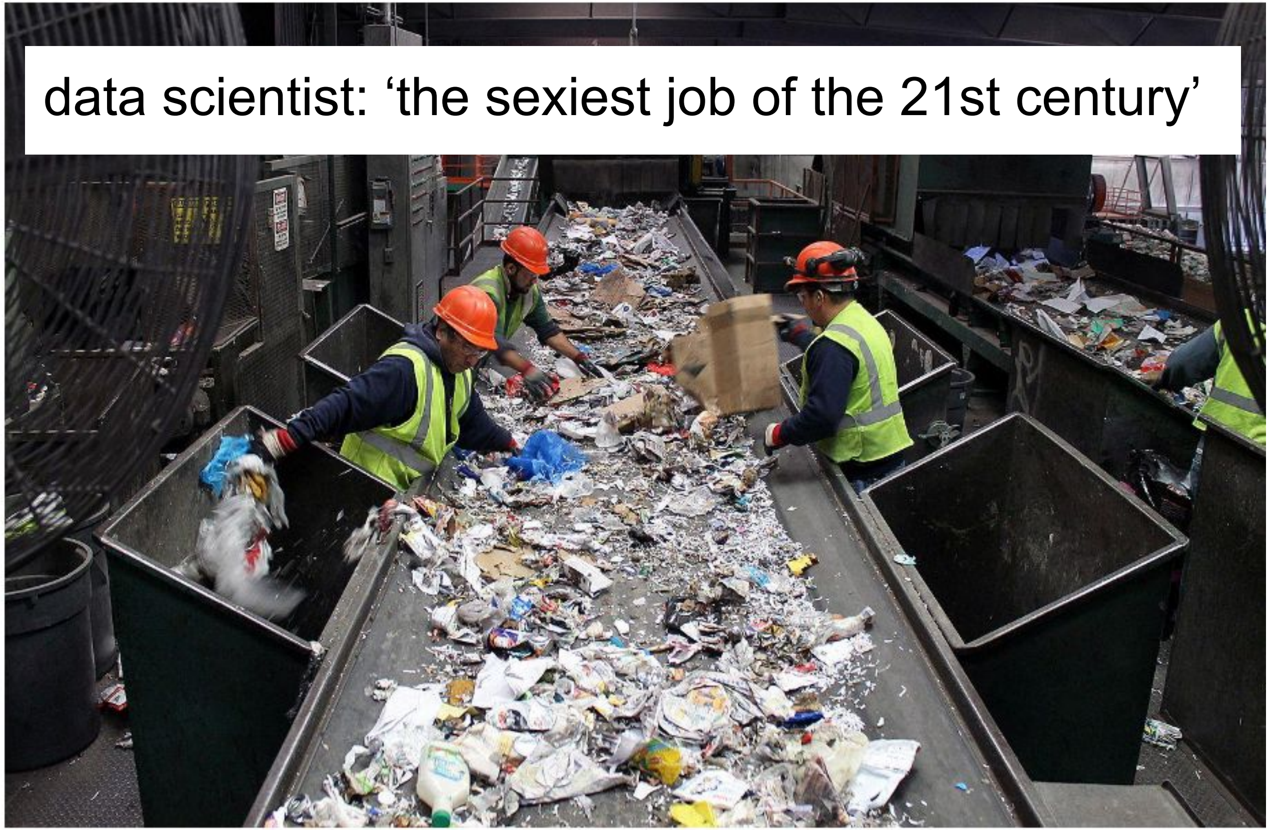
Plan

1. Projects
2. Recap
    a. Optimal Bayes Classifier
    b. Quadratic Discriminant Analysis
    c. Linear Discriminant Analysis
    d. Logistic Regression
3. Demo notebooks
4. Application task: gene expression prediction

Projects: rules

- Teams allowed: 1 - 3 people
- Reporting:
  - Presentation with 7-10 slides summarizing the results of the project and personal contributions
  - Written report 15-20 pages (10-15 for solo projects) of typewritten text shall cover:
    - Problem statement
    - State-of-the-art review
    - Analysis and comparison of applied approaches and solutions
    - Conclusion and references
- 2 tracks: a ‘data scientist simulator’ and a ‘custom project’

Projects: track 1



Projects: track 1 - data scientist simulator

- Select a dataset out of the proposed options and examine it with machine learning techniques studied during the course (some extra techniques are welcome as complements)
- Datasets (will be provided) & problems examples:
  - Sentiment analysis in twitter
  - Facial emotion annotation
  - Grocery shopping baskets analytics
  - Web banners clicks analytics
  - Music annotation
  - A collection of stories
  - a few more...

Projects: track 1 - application form

- To participate in this track, apply with the following form:
  - Selected dataset
  - Specify a meaningful subject of investigation (e.g. a sensitivity analysis w.r.t. some identified parameters, building a classification/regression model for the determined target values, feature engineering and selection etc.)
  - Preliminary ideas regarding investigation approaches
  - Team members, team name

Projects: track 2 - custom project application form

- If you have an interesting proposal or a joint project with another course, then describe it in the following form:
  - Project name
  - Description (1 paragraph) and the main goal
  - Motivation: what makes it interesting/practical
  - Where would data come from?
  - Team members, team name
  - References to relevant papers and datasets

Projects: assessment criteria

10% - General literacy and style of the report

20% - Analytical/scientific methods and approaches

45% - Depth of the subject understanding

25% - Presentation style and Q&A

Unconvincing personal contributions will be penalized individually.

Optimal Bayes Classifier

- Expected prediction error

$$R(C) = \mathbb{E}_{x,y}[\mathcal{L}(y, C(x))] = \mathbb{E}_x\left[\sum_{k=1}^{K} \mathcal{L}(y_k, C(x))Pr(y_k|x)\right]$$

- Optimal Bayes Classifier

$$C^*(x) = \operatorname{argmin} R(C) = \operatorname*{argmin}_{\hat{y}\in Y}\left[\sum_{k=1}^{K} \mathcal{L}(y_k, \hat{y})Pr(y_k|x)\right]$$

- Bayes-optimal Decision Boundary (between 2 classes)

$$\forall y \ \mathcal{L}(y, y) = 0$$
$$\mathcal{L}(y_+, y_-)Pr(y_+|x) = \mathcal{L}(y_-, y_+)Pr(y_-|x)$$

Quadratic Discriminant Analysis

- A special case of bayesian classification
- Assumes that classes have n-dimensional gaussian distributions

$$p(x|y) = \mathcal{N}(x, \mu_y, \Sigma_y) = \frac{1}{\sqrt{(2\pi)^n \det\Sigma_y}} \mathrm{Exp}\Big(-\frac{1}{2}(x-\mu_y)^T \Sigma_y^{-1}(x-\mu_y)\Big)$$

$$\ln p(x|y) = -\frac{n}{2}\ln 2\pi - \frac{1}{2}\ln\,\det\Sigma_y - \frac{1}{2}(x-\mu_y)^T \Sigma_y^{-1}(x-\mu_y)$$

- Optimal bayes classifier with this assumption induces a decision boundary in the quadratic form of x:

$$\mathcal{L}(y_+, y_-)Pr(y_+|x) = \mathcal{L}(y_-, y_+)Pr(y_-|x)$$

$$\mathcal{L}(y_+, y_-)Pr(y_+)p(x|y_+) = \mathcal{L}(y_-, y_+)Pr(y_-)p(x|y_-)$$

$$\ln p(x|y_+) - \ln p(x|y_-) = \ln\frac{\mathcal{L}(y_-, y_+)Pr(y_-)}{\mathcal{L}(y_+, y_-)Pr(y_+)} = const(x)$$

Quadratic Discriminant Analysis

- Has a closed-form solution expressed via MLE of distributions parameters

$$C^*(x) = \underset{\hat{y} \in Y}{\operatorname{argmin}} \left[\sum_{k=1}^{K} \mathcal{L}(y_k, \hat{y}) Pr(y_k) \mathcal{N}(x, \mu_{y_k}, \Sigma_{y_k}) \right]$$

$$\widehat{Pr}(y_k) = \frac{N_k}{N}$$

$$\hat{\mu}_k = \frac{1}{N_k} \sum_{y(x_i)=k} x_i$$

$$\widehat{\Sigma_k} = \frac{1}{N_k - 1} \sum_{y(x_i)=k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T$$

- Disadvantages:
  - Linearly dependent features make covariance matrix be non-invertible
  - Sensitive to gross outliers
  - Not applicable if amount of points in a class is less than the number of features

Linear Discriminant Analysis

- Assumes that covariation matrices for classes are equal

$$\forall y \; \Sigma_y \equiv \Sigma$$

- Increases numerical stability of the covariation matrix estimation

$$\widehat{\Sigma} = \frac{1}{N - |Y|} \sum_{i=1}^{N} (x_i - \hat{\mu}_{y(x_i)})(x_i - \hat{\mu}_{y(x_i)})^T$$

- Simplifies a decision boundary to the linear form of x

$$const(x) = \ln p(x|y_+) - \ln p(x|y_-) = x^T \Sigma^{-1}(\mu_+ - \mu_-) - \frac{1}{2}\mu_+^T \Sigma^{-1} \mu_+ + \frac{1}{2}\mu_-^T \Sigma^{-1} \mu_-$$

Logistic regression

- Assumes that classes have distributions from the exponential family

$$p(x|y) = h(x)g(\theta_y)\mathrm{Exp}(\langle \eta(\theta_y), x \rangle)$$

- Optimal bayes decision boundary under this assumption has a linear form of x

$$const(x) = \ln p(x|y_+) - \ln p(x|y_-) = \langle w, x \rangle$$
$$w = \eta(\theta_+) - \eta(\theta_-)$$

- Posterior probabilities can be explicitly expressed

$$\frac{P(y_+|x)}{P(y_-|x)} = \mathrm{Exp}(\langle w, x \rangle + w_0) \quad \Longrightarrow \quad P(y_\pm|x) = \sigma(\pm(\langle w, x \rangle + w_0))$$

$$P(y_+|x) + P(y_-|x) = 1 \qquad \sigma(z) = \frac{1}{1 + \mathrm{Exp}(-z)}$$

- MLE of w is equivalent to the prediction error minimization with logistic loss

$$\mathcal{L}(y_\pm, C(x)) = \log_2\left(1 + e^{\mp C(x)}\right)$$

Logistic regression regularization

- Log-likelihood

$$L(w, w0, X_{i=1}^{N}) = \log_2 \prod_{i=1}^{N} p(x_i, y(x_i)) = \sum_{i=1}^{N} \left[\log_2 P(y(x_i)|x_i) + \log_2 p(x_i) \right] =$$

$$= \sum_{i=1}^{N} \log_2 \sigma(\pm_i(\langle w, x_i \rangle + w_0)) + const(w, w_0) \to \max_{w,w0}$$

- L1

$$\sum_{i=1}^{N} \log_2 \sigma(\pm_i(\langle w, x_i \rangle + w_0)) \boxed{- \lambda \|w\|_1}$$

- L2

$$\sum_{i=1}^{N} \log_2 \sigma(\pm_i(\langle w, x_i \rangle + w_0)) \boxed{- \lambda \|w\|_2^2}$$

Logistic Regression or LDA?

- LDA has a closed-form solution
- Unlike LDA assumes a wider class of distributions
- LDA has $n|Y|+n(n+1)/2$ parameters, which can be redundant, whereas Logistic Regression has only n
- Logistic regression provides explicit posterior probabilities
- If gaussian assumption is correct, then Logistic Regression asymptotically needs 30% more data to grade up to LDA in terms of error rate
- Strong assumptions on distribution in LDA can be beneficial for semi-supervised learning

Demo notebooks

https://drive.google.com/file/d/0B8-5d8BzFWHgWW1jbUtTb2Q3RGc/view?usp=sharing

Application task

https://inclass.kaggle.com/c/gene-expression-prediction

Notebook with a template:

https://drive.google.com/file/d/0B8-5d8BzFWHgOFNKaVZucVZIR0U/view?usp=sharing

Report here:

https://goo.gl/forms/C0POLQTfA7BOXz2E2