Statistical Learning for Data Science

MSDS534

Lecture 1: Introduction and Review[1]

Department of Statistics
Rutgers University

Acknowledgement

- Some of the figures in this presentation are taken from *An Introduction to Statistical Learning, with applications in R* (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani.
- Some of the figures in this presentation are taken from *Elements of Statistical Learning* (Springer, 2009) with permission from the authors: T. Hastie, R. Tibshirani and J. Friedman.
- Some of the Lab codes in this presentation are taken from *An Introduction to Statistical Learning, with applications in R* (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani.
- Some of these slides are from lectures by Prof. Zijian Guo, and Prof. Lingzhou Xue at Penn State.

Data Sources

- UCI Machine Learning Repository: `https://archive.ics.uci.edu/ml/index.php`
- ESL: `https://web.stanford.edu/~hastie/ElemStatLearn/`
- ISL: `http://faculty.marshall.usc.edu/gareth-james/ISL/`

Reading Assignments

- ESL: 1, §2.1–§2.6, §3.1–§3.6, §4.1 – §4.4.
- ISL: 1–4.

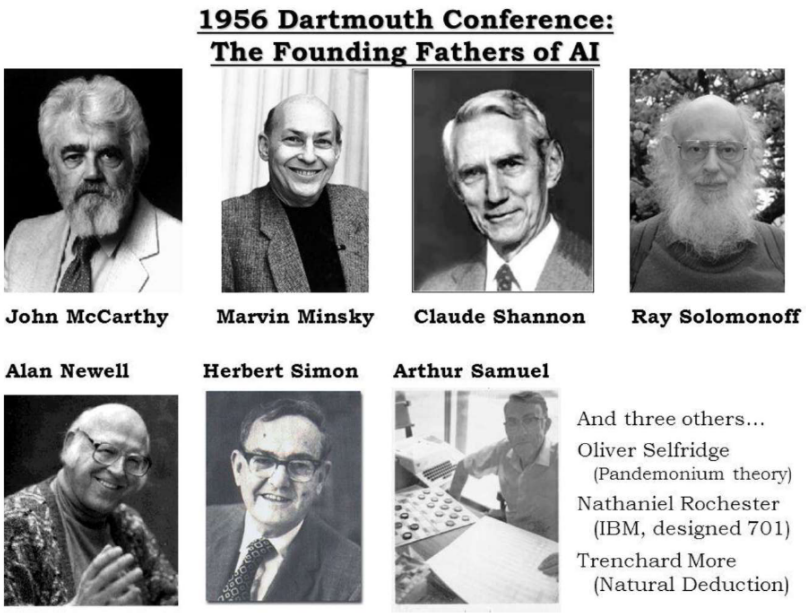
Outline

Dartmouth summer workshop on AI, 1956

Trenchard More, John McCarthy, Marvin Minsky, Oliver Selfridge, and Ray Solomonoff reunited at AI@50 conf.

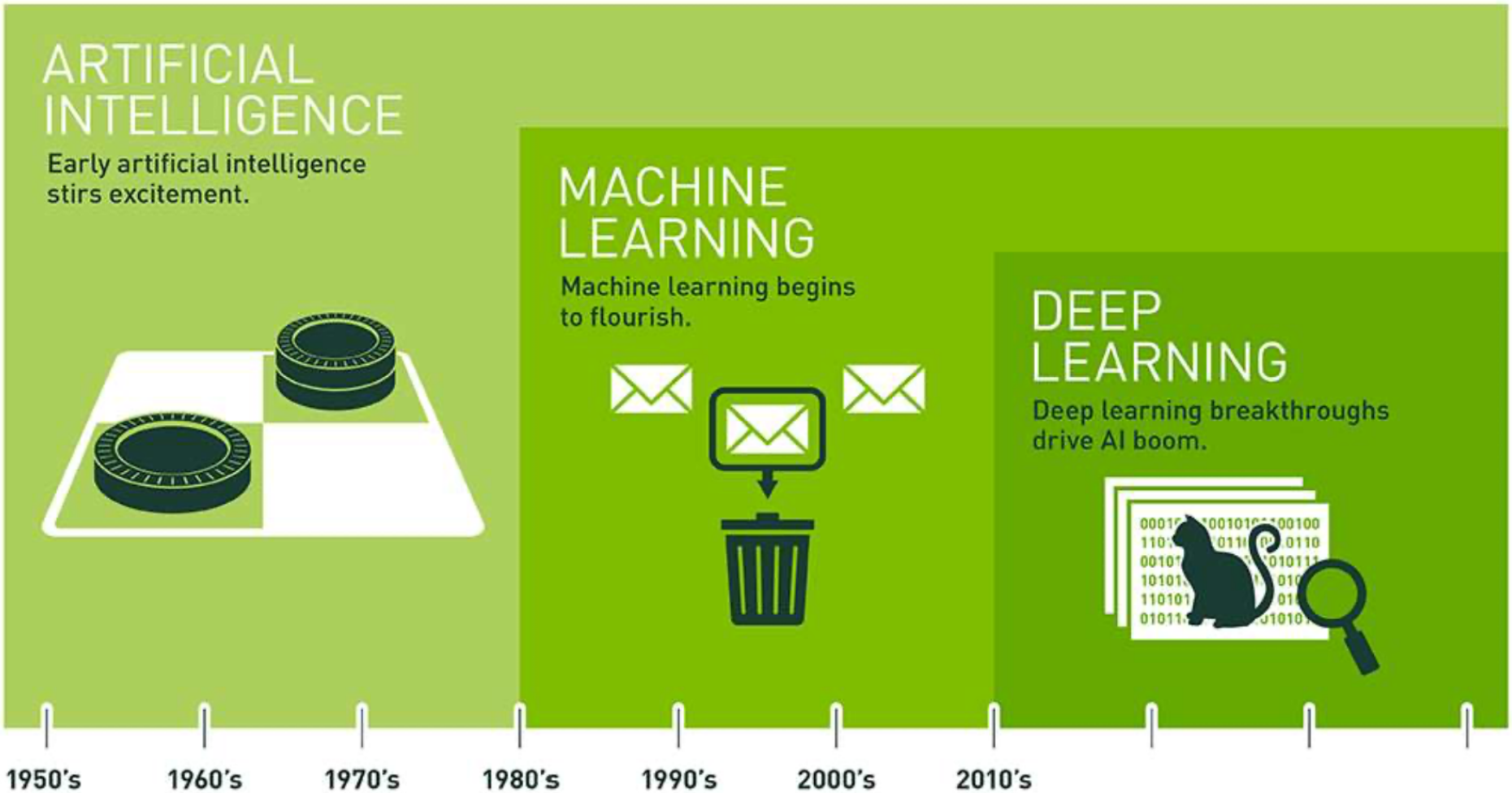
Legends

Founding Fathers of AI



- John McCarthy (1927-2011)
  - Coined the term “Artificial Intelligence” (1956)
  - ACM Turing Award (1971), National Medal of Science (1990)
- Marvin Minsky (1927-2016)
  - Co-founder of MIT CSAIL
  - ACM Turing Award (1969)
- Claude Shannon (1916-2001)
  - Father of Information Theory
  - National Medal of Science (1966), Shannon Award (1972)
- Ray Solomonoff (1926-2009)
  - Father of Algorithmic Probability & Algorithmic Information Theory
  - Kolmogorov Award (2003)
- Allen Newell (1927-1992)
  - ACM Turing Award (1975), National Medal of Science (1992)
- Herbert Simon (1916-2001)
  - ACM Turing Award (1975), National Medal of Science (1986)
  - Nobel Prize in Economics (1978)
- Arthur Samuel (1901-1990)
  - Coined the term “Machine Learning” (1959)
  - Worked with Donald Knuth on the TeX project

AI, ML, DL



Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

AI, ML, IA, II[2]

- Artificial intelligence (coined by McCarthy)
  - The aspiration: human-imitative AI, the high-level or cognitive capability of humans to reason and to think.
  - Current development: engineering (pattern recognition, movement control...), statistics (finding pattern in data, making predictions, tests of hypothesis, decisions...).
  - Wiener coined cybernetics: a vision closely tied to operations research, statistics, pattern recognition, information theory, and control theory.
  - Wiener's agenda dominates under the banner of McCarthy's terminology.
- Machine learning
  - algorithmic field that blends ideas from statistics, computer science and many other disciplines
  - design algorithms that process data, make predictions, and help make decisions
- Intelligent augmentation (IA). Computation and data are used to create services that augment human intelligence and creativity.
  - E.g. search engine, machine translation.
  - Will want computers to trigger new levels of human creativity.
- Intelligent infrastructure (II). A web of computation, data, and physical entities that makes human environments more supportive, interesting, and safe.

[2] Michael I. Jordan. `https://hdsr.mitpress.mit.edu/pub/wot7mkc1/release/8`

Hypothetical reasoning[3]

- People are smarter when they correctly use hypothetical reasoning. They access truths beyond those which can be inferred by simple direct measurement. They explore counterfactual situations they can imagine but not inhabit.
- Statistical hypothesis testing, the cornerstone of statistical inference, is a prime example of hypothetical reasoning.
- Formal statistical models offer another powerful mode of hypothetical reasoning with data.
- Two fields of special relevance.
  - Causal inference. Uses counterfactual reasoning to determine if the data we see might have been different if a certain hypothesized causal factor were silent.
  - Robust inference. Uses a hybrid of worst-case reasoning and statistical modeling to envision consequences of such hypothetical contamination and protect against them.
- True intelligence requires lots of (data-free) hypothetical reasoning about suspected causes, backed up by empirical checks.

[3] David Donoho. `https://hdsr.mitpress.mit.edu/pub/rim3pvdw/release/5`

Role of Statistics[4]

- Robustness. Statistical reasoning and tools will be important.
  - Can we have "good enough" performance 99% of the time?
  - Can we be confident in our predictions?
  - How confident are our predictions?
- Validity of algorithmic inference. A deeper understanding of the data sources and the computations applied will be essential. Statistical quantification of uncertainty.
- Fairness.
  - Biased data collection yields biased results.
  - Statistical calculus of uncertainty, robustness, conditioning, (sub-)population quantities, and prediction errors are important.
- Privacy.
  - A sophisticated literature of algorithmic techniques under privacy constraints is growing.
  - More carefully integrated statistical reasoning is likely to yield tremendous benefits.

[4] Candès et al. `https://hdsr.mitpress.mit.edu/pub/djb16hzl/release/4`

Outline

Regression problem

- Quantitative generic output variable y.
- Generic input vector $\boldsymbol{x} = (x_1, \ldots, x_p)'$.
- Use the regression function $f(\boldsymbol{x})$ to predict y.
- Model the possible relationship between $(\boldsymbol{x}, y)$ by a joint probability distribution.
- Take the loss function as

$$L[y, f(\boldsymbol{x})] = \mathbb{E}_{\boldsymbol{x},y}[y - f(\boldsymbol{x})]^2,$$

  then the optimal regression function is

$$f(\boldsymbol{x}) = \mathbb{E}_{y|\boldsymbol{x}}(y|\boldsymbol{x}).$$

- Typically we have a set of training data $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)$, from which we want to learn the regression function $f(\boldsymbol{x})$, or in statistical language, to estimate $f(\boldsymbol{x})$.

Linear regression models

- Linear regression assumes that $f(\boldsymbol{x})$ takes the form

$$f(\boldsymbol{x}) = \beta_0 + \sum_{j=1}^{p} x_j \beta_j.$$

- Terminology.
  - y: response, output, outcome, dependent variable
  - x_j: covariate, predictor, input, explanatory variable, feature, attribute, independent variable
  - β_j: parameter, regression coefficient
- The input x_j can come from different sources:
  - quantitative inputs;
  - transformations of quantitative inputs: log, square-root or square;
  - basis expansions: polynomial, Fourier expansion, or splines;
  - dummy variables derived from a discrete input;
  - interactions, e.g., $x_3 = x_1 \cdot x_2$.
- Linear means f is a linear function of β_j, $0 \leq j \leq p$.
  - $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_2^2$.

LSE: multiple linear regression

- Observations: $(y_i, x_{i1}, \ldots, x_{ip})$, $1 \le i \le N$.
- Define

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}, \quad \boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N1} & x_{N2} & \cdots & x_{Np} \end{pmatrix}, \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_N \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$$

- Let $\boldsymbol{x}_i = (1, x_{i1}, x_{i2}, \ldots, x_{ip})'$ (abuse of notation), then for each individual observation

$$y_i = \boldsymbol{x}_i' \boldsymbol{\beta} + \epsilon_i, \quad 1 \le i \le N.$$

- The overall linear model can be represented as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}.$$

  - $\boldsymbol{X}$ is called design matrix or model matrix.
- Assume $N > p + 1$ and $\boldsymbol{X}$ has full rank $p + 1$.

Least squares

- Minimize the sum of squared errors

$$\min_{\beta_0,\beta_1,\ldots,\beta_p} \sum_{i=1}^{N} [y_i - (\beta_0 + \beta_1 x_{i1} + \beta_1 x_{i2} + \cdots + \beta_p x_{ip})]^2.$$

- In matrix notation

$$\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2.$$

- Normal equation:

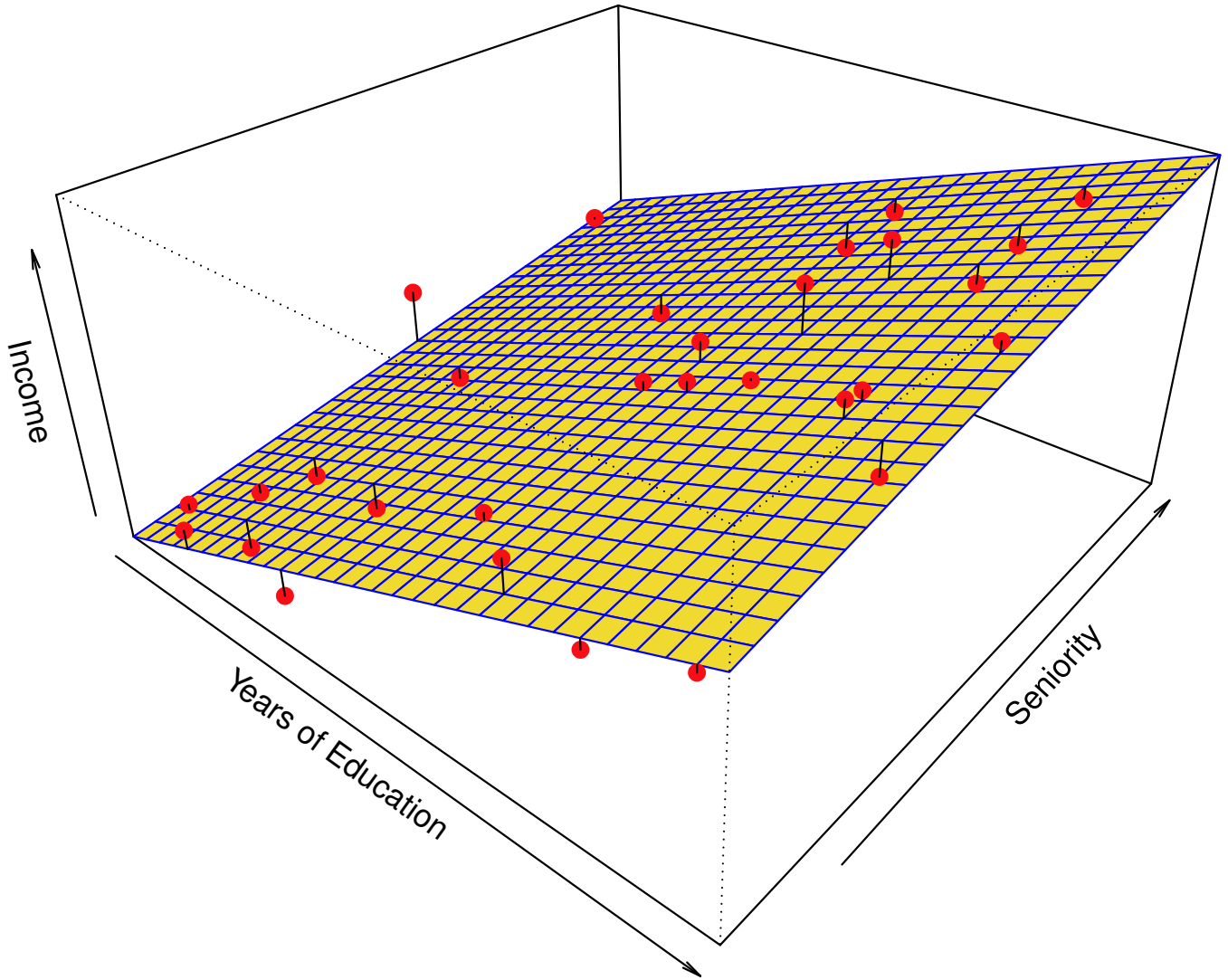
$$\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{y}.$$

- Least squares estimator:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}.$$

- Hat matrix: $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$.
- $\hat{y}_i = \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}$: fitted values. Vector form: $\hat{\boldsymbol{y}} = \boldsymbol{H}\boldsymbol{y}$.
- $\hat{\epsilon}_i = y_i - \hat{y}_i$: residuals.

Least squares: Income vs Education and Seniority[5]



[5] Figure 2.4 of ISL.

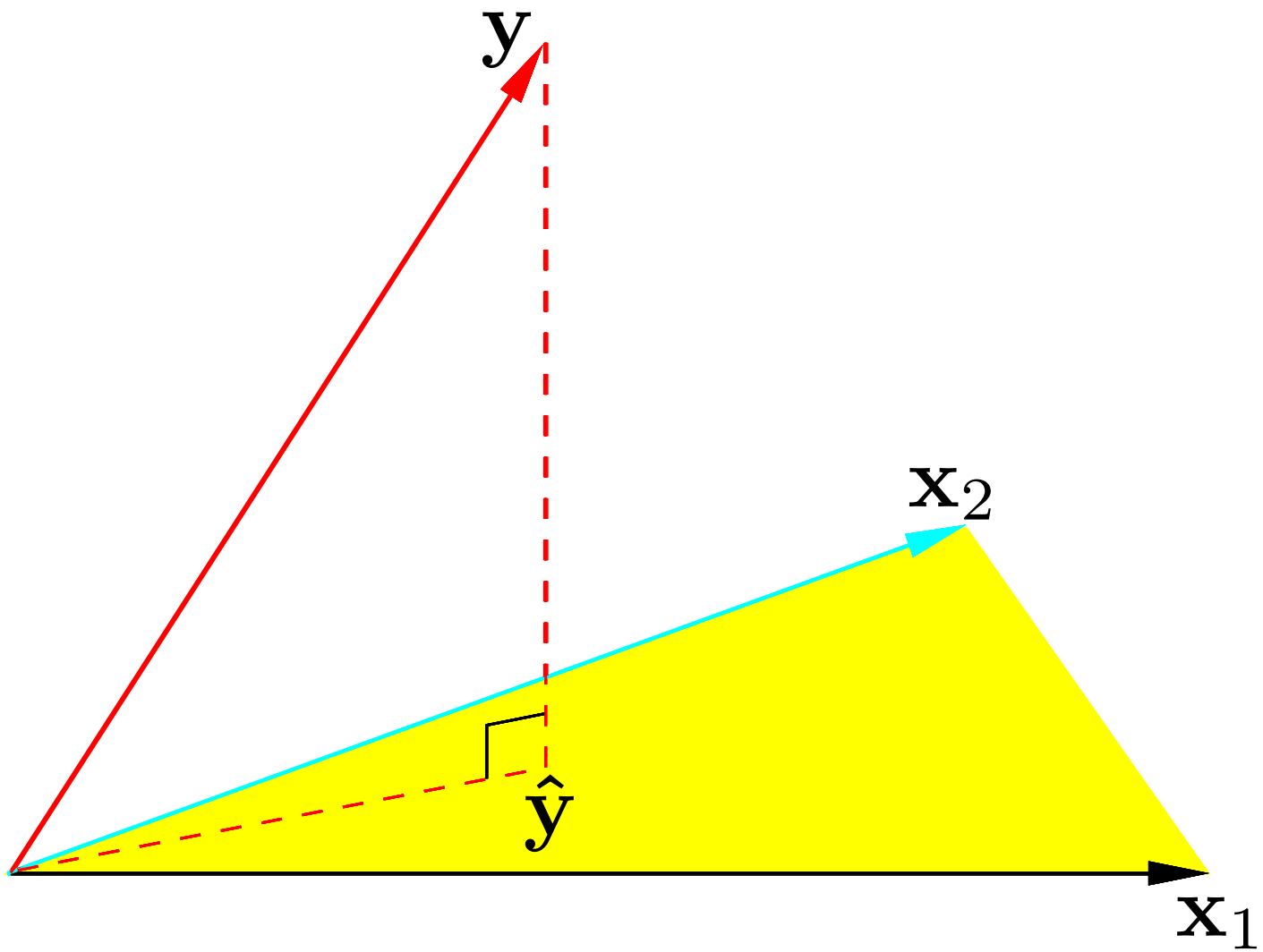
Orthogonal projection

- Let $\boldsymbol{x}^0, \boldsymbol{x}^1, \boldsymbol{x}^2, \ldots, \boldsymbol{x}^p$ be the $p+1$ column vectors of $\boldsymbol{X}$.
- The column span of $\boldsymbol{X}$, denoted by $\mathrm{col}(\boldsymbol{X})$ is the linear subspace consisting of all linear combinations of the form

$$b_0\boldsymbol{x}^0 + b_1\boldsymbol{x}^1 + b_2\boldsymbol{x}^2 + \cdots + b_p\boldsymbol{x}^p, \quad b_0, b_1, \ldots, b_p \in \mathbb{R}.$$

- Least squares method finds a $\hat{\boldsymbol{\beta}}$ such that $\boldsymbol{X}\hat{\boldsymbol{\beta}}$ is the projection of $\boldsymbol{y}$ onto the subspace $\mathrm{col}(\boldsymbol{X})$.
- Let $\boldsymbol{H}$ be the projection matrix onto $\mathrm{col}(\boldsymbol{X})$.
  - Each column of $\boldsymbol{H}$ belongs to $\mathrm{col}(\boldsymbol{X})$, i.e. $\boldsymbol{H} = \boldsymbol{X}\boldsymbol{C}'$ for some $n \times (p+1)$ matrix $\boldsymbol{C}$.
  - $\boldsymbol{H}\boldsymbol{X} = \boldsymbol{X}$.
  - $\boldsymbol{H} = \boldsymbol{H}'$, orthogonal projection.
- $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$.

Orthogonal projection[6]



[6]Figure 3.2 of ESL.

Variance estimator

- Recall the linear model is

$$y_i = \boldsymbol{x}_i' \boldsymbol{\beta} + \epsilon_i, \quad 1 \le i \le N.$$

- Assume $\epsilon_1, \epsilon_2, \ldots, \epsilon_N$ are independent and identically distributed (iid) with mean zero and variance σ^2.
- The estimator of σ^2 is given by

$$\hat{\sigma}^2 = \frac{1}{N-p-1} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2.$$

  - Unbiased estimator, i.e. $\mathbb{E}\hat{\sigma}^2 = \sigma^2$.
- The covariance matrix of $\hat{\boldsymbol{\beta}}$ is

$$\mathrm{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^2 (\boldsymbol{X}'\boldsymbol{X})^{-1}.$$

  - It is estimated by $\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2 (\boldsymbol{X}'\boldsymbol{X})^{-1}$.
- If the assumption is stronger: $\epsilon_1, \epsilon_2, \ldots, \epsilon_N$ are iid $N(0, \sigma^2)$, then $\hat{\boldsymbol{\beta}}$ is the maximum likelihood estimator (MLE) of $\boldsymbol{\beta}$, and the MLE of σ^2 is

$$\tilde{\sigma}^2 = \frac{1}{N} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

Outline

Ridge regression

- The ridge regression solves the optimization problem

$$\hat{\boldsymbol{\beta}}^{\text{ridge}} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

  for some $\lambda \geq 0$.

- An equivalent form is

$$\hat{\boldsymbol{\beta}}^{\text{ridge}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2$$

$$\text{subject to} \quad \sum_{j=1}^{p} \beta_j^2 \leq t,$$

  for some $t \geq 0$.

Ridge regression

- Here $\lambda \geq 0$ is a complexity parameter that controls the amount of shrinkage: the larger the value of λ, the greater the amount of shrinkage. The coefficients are shrunk toward zero (and each other).
- The ridge solutions are not equivariant under scaling of the inputs. Usually standardize the inputs before solving the optimization problem.
- The intercept β_0 has been left out of the penalty term.
- Equivalent problem:
  - Center each $\boldsymbol{x}^j$: $x_{ij}^{(c)} := x_{ij} - \bar{x}_j$.
  - Equivalent problem: estimate β_0 by $\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$, and then solve

$$\min_{\beta_1,\ldots,\beta_p} \left\{ \sum_{i=1}^{N} \left(y_i - \bar{y} - \sum_{j=1}^{p} x_{ij}^{(c)} \beta_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}.$$

  - By an abuse of notation, still use x_{ij} for $x_{ij}^{(c)}$.

Ridge regression

- Pre-step.
  - **(i)** The output vector $\boldsymbol{y}$ is centered, that is, $\sum_{i=1}^{N} y_i = 0$;
  - **(ii)** Each predictor $\boldsymbol{x}^j$, $1 \leq j \leq p$ is normalized, *i.e.*

$$\sum_{i=1}^{N} x_{ij} = 0 \text{ and } \sum_{i=1}^{N} x_{ij}^2 = 1, \qquad \forall\, 1 \leq j \leq p;$$

  - **(iii)** $\hat{\beta}_0 = \bar{y}$.
  - **(iv)** The input matrix $\boldsymbol{X}$ has p (rather than $p+1$) columns;
- Solve the problem (here $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$)

$$\hat{\boldsymbol{\beta}}^{\text{ridge}} = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \left\{ \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_2^2 \right\}.$$

- Ridge regression has a closed form solution

$$\hat{\boldsymbol{\beta}}^{\text{ridge}} = \left(\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I}\right)^{-1} \boldsymbol{X}'\boldsymbol{y}.$$

- Compare with the LSE: $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$.

Lasso

- The LASSO solves the optimization problem

$$\hat{\boldsymbol{\beta}}^{\text{lasso}} = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

  for some $\lambda \geq 0$.

- An equivalent Lagrangian form is

$$\hat{\boldsymbol{\beta}}^{\text{lasso}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2$$
$$\text{subject to} \quad \sum_{j=1}^{p} |\beta_j| \leq t,$$

  for some $t \geq 0$.

Solve lasso

- Pre-step.
  - **(i)** The output vector $\boldsymbol{y}$ is centered, that is, $\sum_{i=1}^{N} y_i = 0$;
  - **(ii)** Each predictor $\boldsymbol{x}_{\cdot j}$, $1 \leq j \leq p$ is normalized, *i.e.*

$$\sum_{i=1}^{N} x_{ij} = 0 \text{ and } \sum_{i=1}^{N} x_{ij}^2 = 1, \qquad \forall\, 1 \leq j \leq p;$$

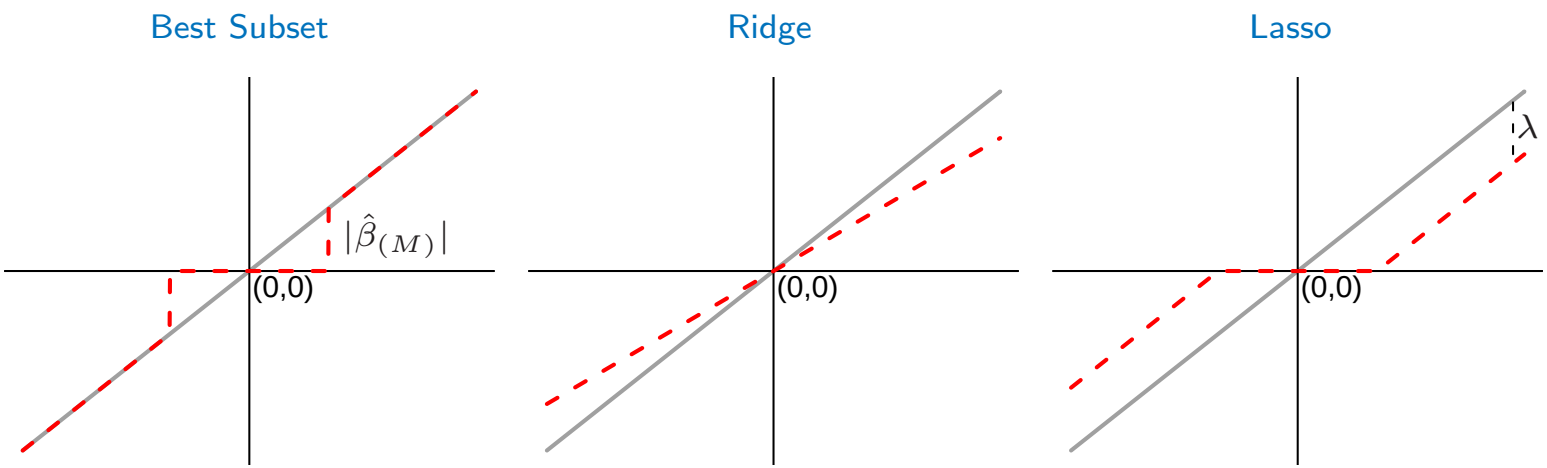
  - **(iii)** $\hat{\beta}_0 = \bar{y}$.
  - **(iv)** The input matrix $\boldsymbol{X}$ has p (rather than $p+1$) columns;
- Solve the problem (here $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$)

$$\begin{aligned}
\hat{\boldsymbol{\beta}}^{\text{lasso}} &= \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2} \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{p} x_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \\
&= \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 \right\}
\end{aligned}$$

Comparison: subset selection, ridge regression and lasso[7]

When the columns of $\boldsymbol{X}$ are orthonormal, the formulas of different methods are given by

Estimator	Formula
Best subset (size M)	$\hat{\beta}_j \cdot I(\lvert\hat{\beta}_j\rvert \geq \lvert\hat{\beta}_{(M)}\rvert)$
Ridge	$\hat{\beta}_j/(1+\lambda)$
Lasso	$\mathrm{sign}(\hat{\beta}_j)(\lvert\hat{\beta}_j\rvert - \lambda)_+$



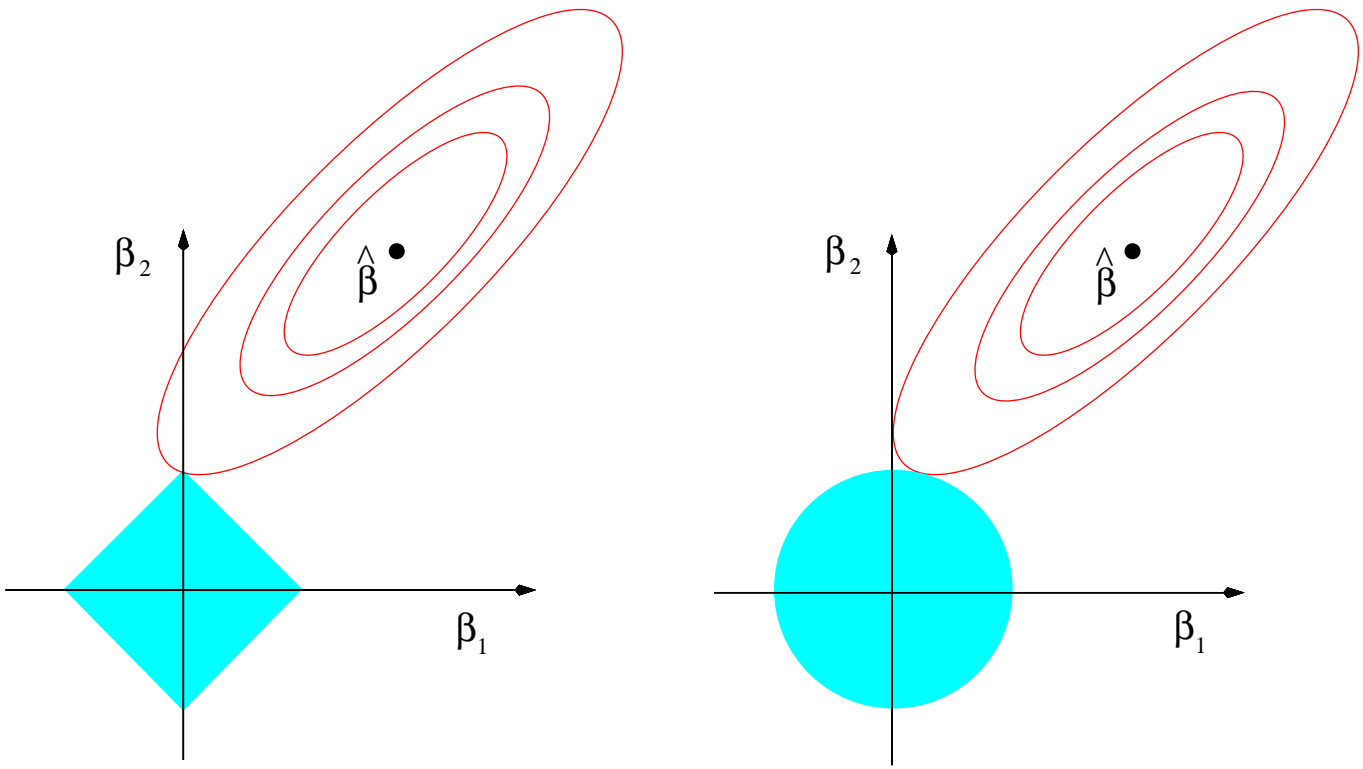
[7] Figure 3.11 of ESL.

Comparison: lasso and ridge regression[8]

$$\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2$$

subject to

$$|\beta_1| + |\beta_2| \leq t \qquad \sqrt{\beta_1^2 + \beta_2^2} \leq t$$



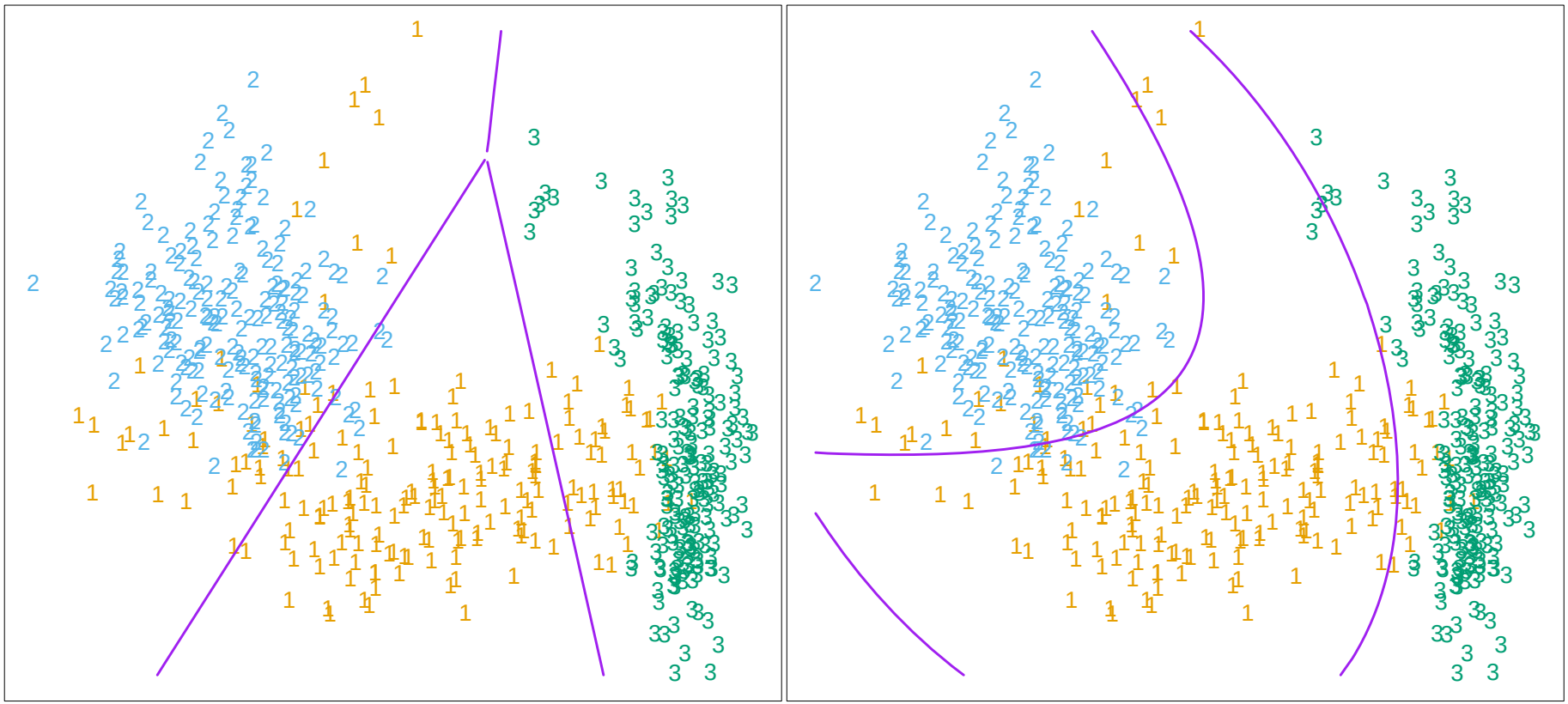
[8] Figure 3.11 of ESL.

Outline

Classification: discriminative boundaries

- Each subject belongs to a class g, and comes with an input $\boldsymbol{x}$. Want to use $\boldsymbol{x}$ to predict g.
- The output g takes value in a discrete set $\mathcal{G}$ with cardinality $|\mathcal{G}| = K$, with each element representing a class. E.g. $\{1, \ldots, K\}$, $\{0, 1, \ldots, K-1\}$, $\{0, 1\}$, $\{-1, 1\}$ etc.
- Divide the input space into a collection of regions labeled by classes. If we have an object $(\boldsymbol{x}, g)$ whose observed input $\boldsymbol{x}$ belong to the region labeled as k, then the unobserved output g is predicted as k.
- Associate each class k with a discriminative function $\delta_k(\cdot)$. For a subject with input $\boldsymbol{x}$, $\hat{g} = \arg\max_k \delta_k(\boldsymbol{x})$.
- The decision boundary between class k and class l is the set of points such that $\delta_k(\boldsymbol{x}) = \delta_l(\boldsymbol{x})$.
- For many methods, either the discriminative function $\delta_k(\boldsymbol{x})$ is linear in $\boldsymbol{x}$, or some monotone transformation of $\delta_k(\boldsymbol{x})$ is linear in $\boldsymbol{x}$. As a result, the decision boundaries are linear in $\boldsymbol{x}$.

Linear and nonlinear decision boundaries[9]



[9] Figure 4.1 of ESL.

Discriminative method: logistic regression

- Consider the binary case, i.e. $g \in \{0, 1\}$, $\boldsymbol{x} \in \mathbb{R}^p$. Logistic regression assumes:

$$\log \frac{P(g = 1|\boldsymbol{x})}{P(g = 0|\boldsymbol{x})} = \beta_0 + \boldsymbol{\beta}'\boldsymbol{x} \quad \text{i.e.}$$

- Equivalently, in terms of the discriminative functions:

$$\delta_0(\boldsymbol{x}) := P(g = 0|\boldsymbol{x}) = \frac{1}{1 + \exp(\beta_0 + \boldsymbol{\beta}'\boldsymbol{x})}$$

$$\delta_1(\boldsymbol{x}) := P(g = 1|\boldsymbol{x}) = \frac{\exp(\beta_0 + \boldsymbol{\beta}'\boldsymbol{x})}{1 + \exp(\beta_0 + \boldsymbol{\beta}'\boldsymbol{x})}.$$

- They are monotone functions of $\beta_0 + \boldsymbol{\beta}'\boldsymbol{x}$, and the decision boundary is

$$\{\boldsymbol{x} : \ \beta_0 + \boldsymbol{\beta}'\boldsymbol{x} = 0\}.$$

Generative method: LDA

- For generative methods, the joint distribution of $(\boldsymbol{x}, g)$ is of interest.
- Bayes Theorem.

$$P(g = k|\boldsymbol{x}) = \frac{P(\boldsymbol{x}|g = k)P(g = k)}{\sum_{l=1}^{K} P(\boldsymbol{x}|g = l)P(g = l)}.$$

- In general, assume the conditional density of $\boldsymbol{x}$ given $g = k$ is given by $f_k(\boldsymbol{x})$. Set $\pi_k = P(g = k)$. We have

$$P(g = k|\boldsymbol{x}) = \frac{\pi_k f_k(\boldsymbol{x})}{\sum_{l=1}^{K} \pi_l f_l(\boldsymbol{x})}.$$

- Assuming that $f_k(\cdot)$ are multivariate normal distributions with different mean vectors, but SAME covariance matrix, then the decision boundary is linear in $\boldsymbol{x}$. Called linear discriminant analysis (LDA).

Generative method: linear discriminant analysis

- Gaussian discriminant analysis assumes

$$f_k(\boldsymbol{x}) = \frac{1}{(2\pi)^{N/2}\det(\Sigma_k)^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)'\Sigma_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)\right\}.$$

- Linear discriminant analysis also assumes $\Sigma_k = \Sigma$ for all $1 \le k \le K$.
- To compare two classes k and l, it suffices to look at the log ratio

$$\log\frac{P(g=k|\boldsymbol{x})}{P(g=l|\boldsymbol{x})} = \log\frac{\pi_k}{\pi_l} - \frac{1}{2}(\boldsymbol{\mu}_k+\boldsymbol{\mu}_l)\Sigma^{-1}(\boldsymbol{\mu}_k-\boldsymbol{\mu}_l) + \boldsymbol{x}'\Sigma^{-1}(\boldsymbol{\mu}_k-\boldsymbol{\mu}_l)$$

- The discriminant functions are given by

$$\delta_k(\boldsymbol{x}) = \log(\pi_k) - \frac{1}{2}\boldsymbol{\mu}_k'\Sigma^{-1}\boldsymbol{\mu}_k + \boldsymbol{\mu}_k'\Sigma^{-1}\boldsymbol{x}, \quad 1 \le k \le K.$$

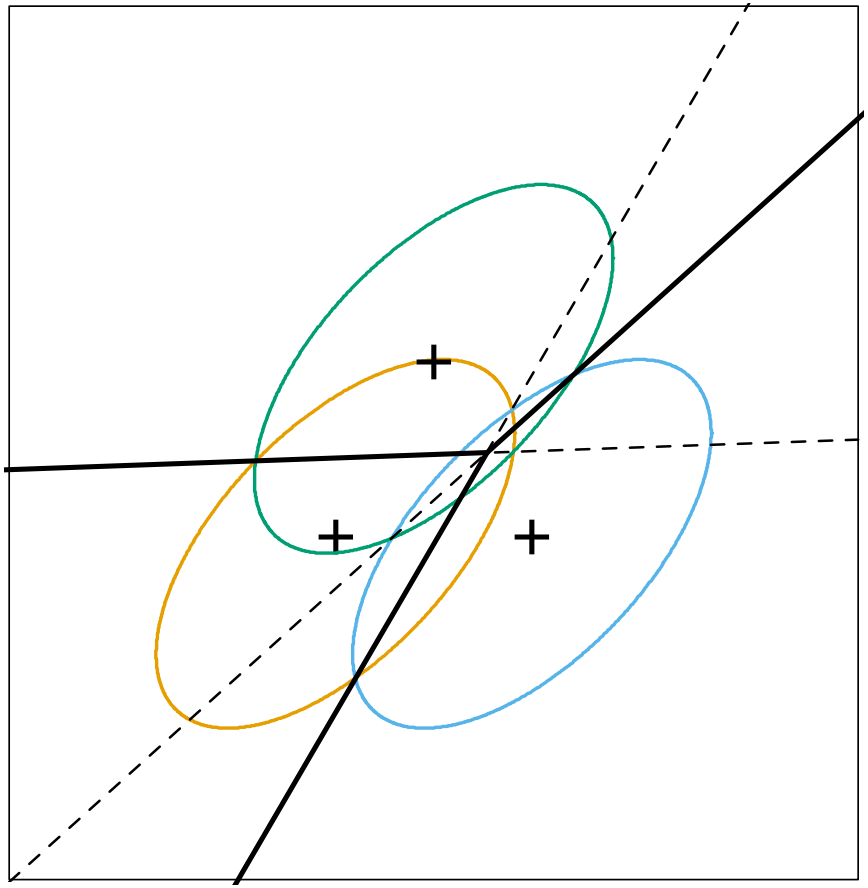
- The parameters $\{\pi_1, \mu_1, \ldots, \pi_K, \mu_K, \Sigma\}$ are estimated by MLE.

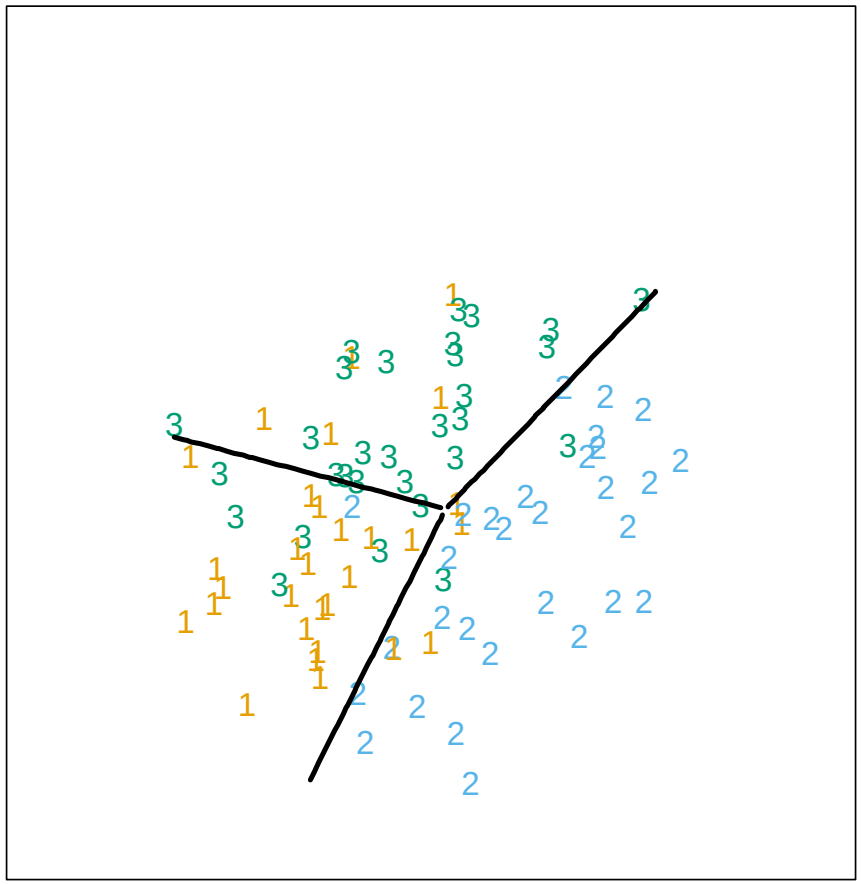
$$\hat{\pi}_k = \frac{1}{N}\sum_{i=1}^{N} I\{g_i = k\}$$

$$\hat{\boldsymbol{\mu}}_k = \frac{\sum_{i=1}^{N}\boldsymbol{x}_i I\{g_i=k\}}{\sum_{i=1}^{N} I\{g_i=k\}}$$

$$\hat{\Sigma} = \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_{g_i})^T(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_{g_i}).$$

LDA[10]





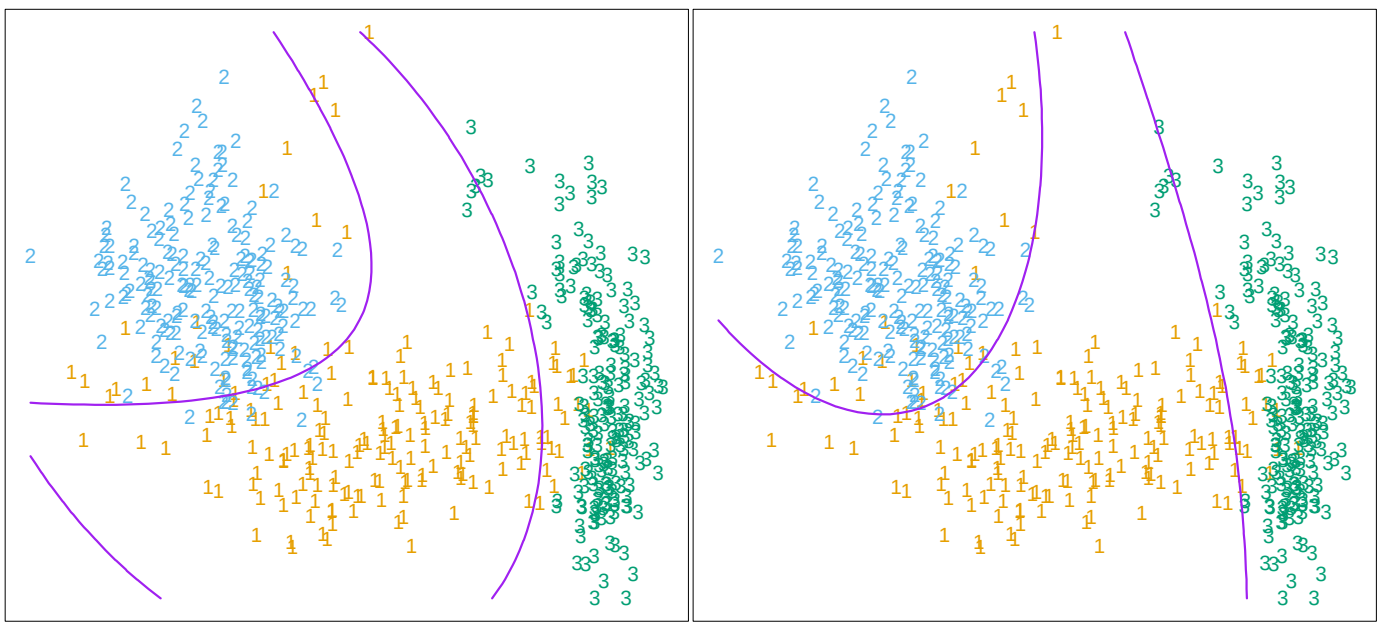
[10] Figure 4.5 of ESL.

QDA and LDA in enlarged input space[11]

- If the Σ_k are not assumed to be equal, we need to use the quadratic discriminant functions

$$\delta_k(\boldsymbol{x}) = \log(\pi_k) - \frac{1}{2}\log[\det(\Sigma_k)] - \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_k)'\Sigma_k^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_k).$$

- Alternatively, we can perform LDA in the enlarged feature space: $\{x_1, \ldots, x_p, x_1^2, \ldots, x_p^2, x_1x_2, \ldots, x_{p-1}x_p\}$.



[11] Figure 4.6 of ESL.