Statistical Learning – week 3.3

Joris Bierkens

Delft University of Technology, The Netherlands

29 February 2024

Outline

1. Maximum Likelihood Estimation
2. Classification
   - Linear regression
   - Decision boundaries
   - Quadratic and Linear Discriminant Analysis
   - Logistic regression

Recap learning objectives week 3.2

- Parametric models (G)
- The linear regression model and the least squares estimator
- Linear regression in the over-parametrized case; Moore-Penrose inverse
- The Gauss-Markov theorem
- Refresher: Lagrange Multipliers (G)
- Regularized linear regression: ridge regression and the LASSO

Maximum Likelihood Estimation

- Consider a family of densities $\{p_{\boldsymbol{\theta}}(\boldsymbol{x}) : \boldsymbol{\theta} \in \Theta\}$.
- Each $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ is a model for the observations $\boldsymbol{x}$.
- Here $\boldsymbol{x}$ may (or may not) consist of multiple independent observations: $\boldsymbol{x} = (x_1, \ldots, x_n)$.
- In frequentist statistics or classical statistics we typically assume there is a 'true' value $\boldsymbol{\theta}_0$.
- We may use the maximum likelihood method to determine an estimator $\hat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}_0$.

$$\hat{\boldsymbol{\theta}} = \arg\max p_{\boldsymbol{\theta}}(\boldsymbol{x}).$$

- Usually easier to work with the log likelihood $\ell(\boldsymbol{\theta}) = \log p_{\boldsymbol{\theta}}(\boldsymbol{x})$.
- uncertainty quantification using confidence intervals.

Example: linear regression

Consider the linear regression model

$$y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \varepsilon_i, \quad \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2) \quad i = 1, \ldots, n.$$

Suppose we wish to estimate $(\boldsymbol{\beta}, \sigma^2)$ from the data using maximum likelihood.

The likelihood is given by

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})^2 / (2\sigma^2)\right).$$

The log likelihood is

$$\ell(\boldsymbol{\beta}) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})^2.$$

Exercise

The maximum likelihood estimator is given by

$$\hat{\boldsymbol{\beta}}_{\mathrm{MLE}} = \hat{\boldsymbol{\beta}}_{\mathrm{OLS}} \quad \text{and} \quad \hat{\sigma}^2_{\mathrm{MLE}} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}_{\mathrm{OLS}})^2 = \frac{1}{n} \mathrm{RSS}(\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}).$$

Maximum Likelihood Estimation: consistency

For observations $\boldsymbol{X}_i$, $i = 1, \ldots, n$ and a model family

$$\{p_{\boldsymbol{\theta}}(\boldsymbol{x}) : \boldsymbol{\theta} \in \Theta\},$$

write the log likelihood as

$$\ell_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log p_{\boldsymbol{\theta}}(\boldsymbol{X}_i).$$

Theorem

Under mild conditions, the MLE is consistent.

This means that, if $\boldsymbol{X}_i \sim p_{\boldsymbol{\theta}_0}(\boldsymbol{x})$, for $i = 1, \ldots, n$, and

$$\hat{\boldsymbol{\theta}}_n = \arg\max_{\boldsymbol{\theta} \in \Theta} \ell_n(\boldsymbol{\theta}),$$

then $\hat{\boldsymbol{\theta}}_n \to \boldsymbol{\theta}_0$ in probability.

Maximum Likelihood Estimation: asymptotic normality

For observations $X_1, \ldots, X_n$ and a family of distributions $\{p_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$, write $\ell_n(\boldsymbol{\theta}) = \sum_{i=1}^n \log p_{\boldsymbol{\theta}}(X_i)$ for the log likelihood and $\hat{\boldsymbol{\theta}}_n = \arg\max_{\boldsymbol{\theta}\in\Theta} \ell_n(\boldsymbol{\theta})$.

Theorem

Under mild conditions,

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\mathcal{I}}_{\boldsymbol{\theta}_0}^{-1}),$$

where the Fisher information is defined as

$$\boldsymbol{\mathcal{I}}_{\boldsymbol{\theta}} = \mathbb{E}_{\boldsymbol{\theta}}\left[(\nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(X))(\nabla \log p_{\boldsymbol{\theta}}(X))^T\right].$$

The Fisher information is often not available exactly:

- We do not know the value of $\boldsymbol{\theta}_0$;
- The required expectation may be difficult to compute.

How to deal with this?

- We may use a $\hat{\boldsymbol{\theta}}_n$ as a plugin-estimator for $\boldsymbol{\theta}_0$.
- The Fisher information may be approximated using the observed Fisher information,

$$\boldsymbol{\mathcal{I}}_{\boldsymbol{\theta}_0} \approx \boldsymbol{\mathcal{I}}_n = -\frac{1}{n}\sum_{i=1}^n (\nabla_{\boldsymbol{\theta}}^2 \log p_{\hat{\boldsymbol{\theta}}}(X_i))\bigg|_{\hat{\boldsymbol{\theta}}_n}.$$

Approximate confidence intervals

We have seen that the MLE $\hat{\boldsymbol{\theta}}_n$ (under some conditions) admits the asymptotic distribution

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathcal{I}_{\boldsymbol{\theta}_0}^{-1}).$$

So if $U_\alpha \subset \mathbb{R}^p$ is such that

$$\mathbb{P}(\mathcal{N}(\mathbf{0}, \boldsymbol{I}_p) \in U_\alpha) = 1 - \alpha,$$

then for

$$\mathcal{C}_\alpha := \{\boldsymbol{\theta} \in \mathbb{R}^p : \sqrt{n}\mathcal{I}_{\boldsymbol{\theta}_0}^{-1/2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \in U_\alpha\},$$

we find that

$$\mathbb{P}(\boldsymbol{\theta}_0 \in \mathcal{C}_\alpha) \approx 1 - \alpha.$$

Linear regression for classification

Indicator response

- Classification: outcomes in a finite set $\mathcal{Y}$.
- Observations $(\boldsymbol{x}_i, y_i)$ for $\boldsymbol{x}_i \in \mathbb{R}^d$ and $y_i \in \mathcal{Y}$.
- Equivalently: observations $(\boldsymbol{x}_i, \boldsymbol{z}_i)$ where $\boldsymbol{z}_i \in \{0,1\}^{\mathcal{Y}}$.

$$\boldsymbol{z}_i(k) = \begin{cases} 1 & (y_i = k), \\ 0 & (y_i \neq k). \end{cases}$$

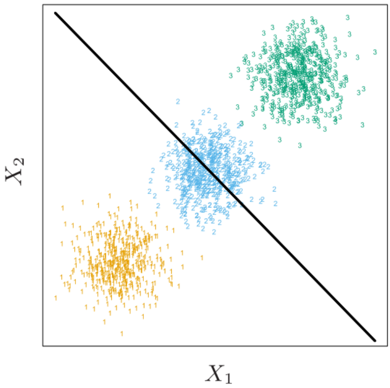
- $|\mathcal{Y}|$ regression problems: for $k \in \mathcal{Y}$,

$$\boldsymbol{z}_i(k) = \boldsymbol{x}_i^T \boldsymbol{\beta}_k + \varepsilon_{i,k}.$$

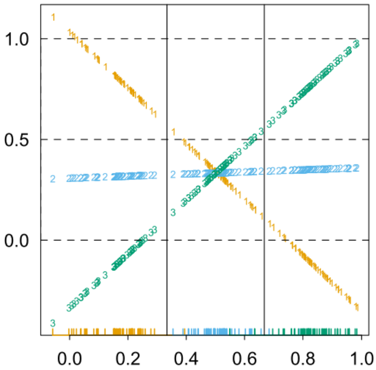
- Classify new input x by finding

$$\hat{y}(\boldsymbol{x}) = \arg\max_{k \in \mathcal{Y}} \boldsymbol{x}^T \boldsymbol{\beta}_k.$$

Masking



(a) A three-cluster model and the linear regression decision boundary



(b) Linear regression slopes along the diagonal

Take away: linear regression is not suitable for classification

Decision boundaries in classification

Recall the generative model for classification: For $\boldsymbol{x} \in \mathbb{R}^d$ and $y \in \mathcal{Y}$,

- class-conditional densities $p(\boldsymbol{x} \mid y)$, and
- prior class probabilities $\pi_k := p(y = k)$.

The posterior class probabilities are given by Bayes rule

$$p(y \mid \boldsymbol{x}) = \frac{p(\boldsymbol{x}, y)}{p(\boldsymbol{x})} = \frac{p(\boldsymbol{x} \mid y)p(y)}{\sum_{y' \in \mathcal{Y}} p(\boldsymbol{x} \mid y')p(y')}.$$

Discriminant function

Discriminant functions $\delta_k : \mathbb{R}^d \to \mathbb{R}$, for $k \in \mathcal{Y}$, are functions satisfying

$$p(y = k \mid \boldsymbol{x}) = \frac{\exp(\delta_k(\boldsymbol{x}))}{\sum_{k'} \exp(\delta_{k'}(\boldsymbol{x}))}.$$

Equivalently $\delta_k(\boldsymbol{x}) = \log p(y = k \mid x) + h(\boldsymbol{x})$ for some function $h(\boldsymbol{x})$.

The decision boundaries are hypersurfaces in $\mathbb{R}^d$ given by

$$\{\boldsymbol{x} \in \mathbb{R}^d : \delta_k(\boldsymbol{x}) = \delta_\ell(\boldsymbol{x})\}, \quad k, \ell \in \mathcal{Y}.$$

Softmax and logistic function

The softmax function

The softmax function $\sigma : \mathbb{R}^{\mathcal{Y}} \to [0,1]^{\mathcal{Y}}$ is given by

$$\sigma(\boldsymbol{a})_k = \frac{\exp(a_k)}{\sum_{k'} \exp(a_{k'})}, \quad k \in \mathcal{Y}.$$

It is invariant under transformation $a'_k = a_k + c$.

Using the softmax function, our posterior class probabilities may be written as

$$p(y = k \mid \boldsymbol{x}) = \frac{\exp(\delta_k(\boldsymbol{x}))}{\sum_{k'} \exp(\delta_{k'}(\boldsymbol{x}))} = \sigma(\delta(\boldsymbol{x}))_k.$$

If $|\mathcal{Y}| = 2$, say $\mathcal{Y} = \{0, 1\}$, this simplifies to

$$p(y = 1 \mid \boldsymbol{x}) = \sigma(\delta(\boldsymbol{x})),$$

where

$$\sigma(a) = \frac{1}{1 + e^{-a}}$$

is the logistic function.

Quadratic Discriminant Analysis

Suppose for $k \in \mathcal{Y}$ the class-conditional probabilities are given by a $\mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ distribution, i.e.

$$p(\boldsymbol{x} \mid y = k) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}_k|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_k)\right).$$

This gives quadratic discriminant functions

$$\delta_k(\boldsymbol{x}) = -\frac{1}{2}\log|\boldsymbol{\Sigma}_k| - \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_k)^T \Sigma_k^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_k) + \log \pi_k.$$

Quadratic discriminant analysis

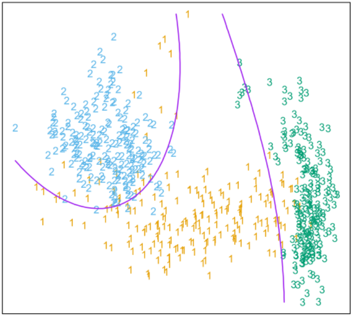
$$\delta_k(x) = -\frac{1}{2}\log|\boldsymbol{\Sigma}_k| - \frac{1}{2}(x-\boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(x-\boldsymbol{\mu}_k) + \log \pi_k$$

- $(\Sigma_k)_{k\in\mathcal{Y}}$, $(\boldsymbol{\mu}_k)_{k\in\mathcal{Y}}$ and $(\pi_k)_{k\in\mathcal{Y}}$ are parameters of the model.
- Unbiased estimators

$$\hat{\pi}_k = n_k/n, \qquad (n_k = \sum_{i=1}^{n} \mathbb{1}_{y_i=k}),$$

$$\hat{\boldsymbol{\mu}}_k = \frac{1}{n_k}\sum_{i:y_i=k} \boldsymbol{x}_j,$$

$$\hat{\boldsymbol{\Sigma}}_k = \frac{1}{n_k-1}\sum_{i:y_i=k} (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_k)(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_k)^T.$$

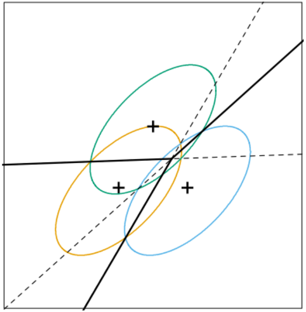


Linear discriminant analysis

Linear discriminant analysis (LDA) arises by assuming that $\boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}$ does not depend on k.

In this case the decision boundaries are hyperplanes determined by the linear discriminant functions

$$\delta_k(x) = \boldsymbol{x}^T \Sigma^{-1} \boldsymbol{\mu}_k - \frac{1}{2} \boldsymbol{\mu}_k^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k + \log \pi_k .$$

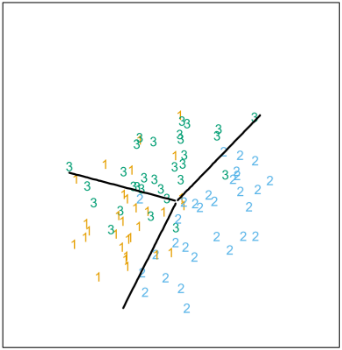


Linear discriminant analysis (LDA)

$$\delta_k(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k - \frac{1}{2} \boldsymbol{\mu}_k^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k + \log \pi_k.$$

- Again, $(\pi_k)_{k \in \mathcal{Y}}$, $(\boldsymbol{\mu}_k)_{k \in \mathcal{Y}}$ and $\boldsymbol{\Sigma}$ are parameters.
- May use the same estimators as before for π_k and $\boldsymbol{\mu}_k$.
- An unbiased estimator for $\boldsymbol{\Sigma}$ is given by

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n - |\mathcal{Y}|} \sum_{k \in \mathcal{Y}} \sum_{i : y_i = k} (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_k)(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_k)^T$$



From LDA to logistic regression

In LDA the posterior class probabilities are fully determined by the discriminant functions of the form $\delta_k(\boldsymbol{x}) = \beta_{k,0} + \boldsymbol{\beta}_k^T \boldsymbol{x}$, i.e.,

$$p(y = k \mid \boldsymbol{x}) = \sigma(\delta(\boldsymbol{x}))_k = \frac{\exp(\delta_k(\boldsymbol{x}))}{\sum_{k'} \exp(\delta_{k'}(\boldsymbol{x}))}.$$

We therefore parametrize using $(\beta_{k,0}, \boldsymbol{\beta}_k)_{k \in \mathcal{Y}}$ instead of $(\boldsymbol{\mu}_k, \pi_k)_{k \in \mathcal{Y}}$ and $\boldsymbol{\Sigma}$.

This is still (slightly) overparametrized: $\sigma(\delta(\boldsymbol{x}))$ is invariant under addition of $\alpha_0 + \boldsymbol{\alpha}^T \boldsymbol{x}$ to each function $\delta_k(x)$. Therefore we set $\delta_K(\boldsymbol{x}) = 0$ for a single $K \in \mathcal{Y}$.

This gives the multinomial logistic regression model

$$\begin{aligned} p(y = k \mid \boldsymbol{x}) &= \frac{\exp(\beta_{k,0} + \boldsymbol{\beta}_k^T \boldsymbol{x})}{1 + \sum_{k \in \mathcal{Y}; k \neq K} \exp(\beta_{k,0} + \boldsymbol{\beta}_k^T \boldsymbol{x})}, \\ p(y = K \mid \boldsymbol{x}) &= \frac{1}{1 + \sum_{k \in \mathcal{Y}; k \neq K} \exp(\beta_{k,0} + \boldsymbol{\beta}_k^T \boldsymbol{x})}. \end{aligned}$$

How to estimate $(\beta_{k,0}, \boldsymbol{\beta}_k)_{k \in \mathcal{Y}, k \neq K}$? Maximum likelihood!

Binary logistic regression

In binary classification we often choose $\mathcal{Y} = \{0, 1\}$.

This gives the binary logistic regression model

$$p_\beta(y = 1 \mid \boldsymbol{x}) = \frac{1}{1 + \exp(-\boldsymbol{x}^T \beta)}.$$

The log likelihood is given by

$$\begin{aligned}
\ell(\boldsymbol{\beta}) &= \sum_{i=1}^{n} y_i \log p_{\boldsymbol{\beta}}(y_i = 1 \mid \boldsymbol{x}_i) + (1 - y_i) \log p_{\boldsymbol{\beta}}(y_i = 0 \mid \boldsymbol{x}_i) \\
&= \sum_{i=1}^{n} y_i \boldsymbol{x}_i^T \boldsymbol{\beta} - \log\left(1 + e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}\right)
\end{aligned}$$

Learning objectives week 3.3

- Discriminant functions in classification
- Quadratic Discriminant Analysis and Linear Discriminant Analysis
- Refresher: Maximum likelihood method (G)
- Logistic regression

Lecture note exercises: 2.1, 2.2, 2.3, 4.1, 4.2, 4.3 and 4.4.

First assignment is now available.