60611 Advanced Statistical Learning
Introduction

Aurélie Labbe (based on D. Larocque course notes)

HEC Montréal
Sciences de la décision

Agenda

Introduction

1. **Introduction**
2. Basic concepts: bias-variance tradeoff
3. Basic concepts: maximum likelihood estimation
4. Model selection
   - Likelihood-based criteria (AIC / BIC)
   - Sample-splitting methods
   - Cross-validation
   - 1 SE rule with cross validation
5. Generalized Linear Models

Let me introduce myself...

- Aurélie LABBE
- 2005: PhD in statistics
- 2005-2009: University Laval, Department of maths and stats
- 2009-2016: University McGill, Département of biostatistics
- 2016- : HEC, Decision Sciences (full prof.)
- Program chair: MSc Data Science and Business Analytics
- 2019: Research Chair FRQ-IVADO in data science

Who are you ?

- Program ? Background ?
  - Business school
  - Engeneering
  - Math/stat
  - Computer science ?
  - Other ?
- Knowledge of R: basic / advanced ?

Course objectives

- Advanced notions of statistical learning:
  - Regularization methods and variable selection
  - Tree based methods and random forests
  - Boosting
  - Survival analysis
  - Prediction intervals
  - Analysis of correlated data
- R is the programming language used in this course

Course pre-requisites

- MATH 60603 Statistical learning

  OR

- MATH 60600 Data Mining + MATH 60602 Analyse multidimensionnelle appliquée
- Knowledge of R is assumed

Learning strategies

- On campus only
- Lectures: slides are provided, as complete as full course notes
- R tutorials: code is provided in the notes
- A .zip file containing all the code and data for the course is provided on Zone Cours.

Communication

- From me to you: always through Zone Cours
- Don't hesitate to contact me by email if you have questions

Evaluations

- **Project** (team work): 25%
  - Already available on Zone cours
  - Case study: road safety pilot study in Montreal
  - Deadline: March, 12th, 23h55
- **Assignment** (individual work): 10%
  - Deadline: April, 16th, 23h55
- **Midterm exam**: 25%
  - Saturday, March 2nd: 13h30 - 16h30
  - You are allowed to bring 3 sheets (double sided) of notes, standard format
- **Final exam**: 40%
  - Tuesday, April 23rd: 13h30 - 16h30
  - You are allowed to bring 5 sheets (double sided) of notes, standard format

Course workload

- 1 credit = 45h of work including lectures
- 3 credits = 7-8 hours of personnal work (outside classes)

Course material (Zone Cours)

- Slides (includes R tutorials): equivalent to full course notes, in slide format.
- **Additionnal book references**:
  - Berk (2008) Statistical learning from a regression perspective
  - Friedman et al. (2009) The elements of statistical learning.
  - James et al. (2013) An introduction to statistical learning,
  - Hastie et al. (2015) Statistical learning with sparsity: the lasso and generalizations.
  - Efron and Hastie (2016) Computer Age Statistical Inference
  - Friedman et al. (2009) The elements of statistical learning.
- Scientific papers cited in the slides: the complete list of references is available on Zone Cours (pdf file)

What you can expect from me...

- Provide course material
- Answer questions by email or by video conference
- Grade exams / homework within 2 weeks

What do I expect from you...

- This is an advanced class...
- Come prepared to class:
  - Lecture material already downloaded from Zone Cours
  - Review lecture material from previous weeks to make sure everything has sunk in
- Participate in class:
  - Attend all lectures
  - Ask questions !!!
  - Discussions

Agenda

Introduction

1. Introduction
2. **Basic concepts: bias-variance tradeoff**
3. Basic concepts: maximum likelihood estimation
4. Model selection
   - Likelihood-based criteria (AIC / BIC)
   - Sample-splitting methods
   - Cross-validation
   - 1 SE rule with cross validation
5. Generalized Linear Models

General framework

- In this course, we consider a target variable (dependent variable, outcome, response) Y
- We also consider many covariates (explanatory variables, predictors) $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)$.
- The goal is to use the covariates to explain and/or predict the target variable.
- The covariates can be of any types:
  - continuous
  - categorical (binary, nominal, ordinal)

General framework: target variable

- The target Y can be of any types:
  - continuous
  - categorical (binary, nominal, ordinal)
- The target Y can be censored (survival data)
- The observations Y could be dependent (e.g. clustered data).
- The target Y can also be a vector (multivariate outcome).

General framework: predicting vs explanatory modeling

- Explanatory modeling uses statistical models for testing causal explanations
- Predictive modeling is the process of applying a statistical model or data mining algorithm to data for the purpose of predicting new or future observations.
- In this course, both situations will be considered although we will spend a lot more time on predictive modeling.

Predicting vs explanatory modeling: basic definitions

- Assume that Y is continuous, and that

$$E(Y|\boldsymbol{X} = \boldsymbol{x}) = g(\boldsymbol{x}),$$

where g is an unknown function.

- Assume we have an estimation of g, say, $\hat{g}$, obtained with any model (linear regression, regression tree, random forest,...).
- The prediction of an observation with $\boldsymbol{X} = \boldsymbol{x}$ is

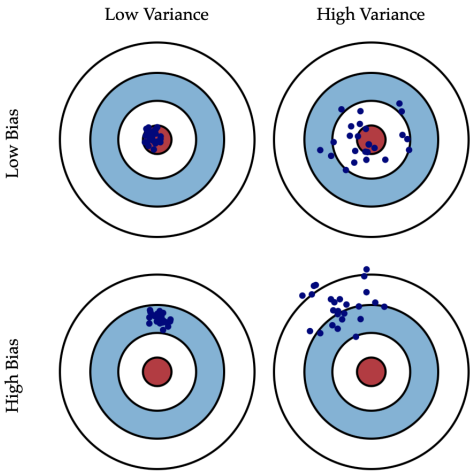
$$\hat{g}(\boldsymbol{x}).$$

- The bias of the prediction is the difference between the true mean and the expected value of the prediction:

$$\text{Bias} = E[\hat{g}(\boldsymbol{x})] - g(\boldsymbol{x})$$

- The variance of the prediction is $Var[\hat{g}(\boldsymbol{x})]$
- The variance of a new observation is simply $Var(Y)$

Illustration of the biais and variance



Source: http://scott.fortmann-roe.com/docs/BiasVariance.html

Predicting vs explanatory modeling: biais-variance tradeoff

- One example of the differences between predicting vs explanatory modeling can be described by the way of an important concept, the expected prediction error (EPE):

$$\text{EPE}(\boldsymbol{x}) = E[(Y - \hat{g}(\boldsymbol{x}))^2]$$

- It is possible to show that the EPE can be decomposed into three terms:

$$\text{EPE}(\boldsymbol{x}) \quad = \quad Var(Y) + \text{Bias}^2 + Var[\hat{g}(\boldsymbol{x})].$$

- The first term, $Var(Y)$, is the error of a new observation that we can never predict. We can not find a model that has a prediction error below that.
- The second and third term are under the control of the analyst and define the bias-variance trade-off.
- **Finding the best predictive model amounts to minimize the sum of the last two terms, that is to find a good compromise between the bias and the variance.**

Predicting vs explanatory modeling: biais-variance tradeoff

- In the case of an explanatory model, we often focus more on having a small bias in order to obtain a good representation of the true underlying model.
- However, using the true model is not always the best predictive model.
- Intuitively, using a biased model which is easier to estimate (with a small variance) may sometimes have a smaller EPE than another model with a smaller bias (even with no bias) if this model has a large variance.
- Let's see a simple example...

Example of biais-variance tradeoff

- Consider two predictors X_1 and X_2 that come from a bivariate normal distribution:
  - Each one has a mean of 0 and a variance of 1.
  - The correlation between them is ρ.
- The true model under which data is generated is:

$$Y = 3X_1 + 0.4X_2 + \epsilon$$

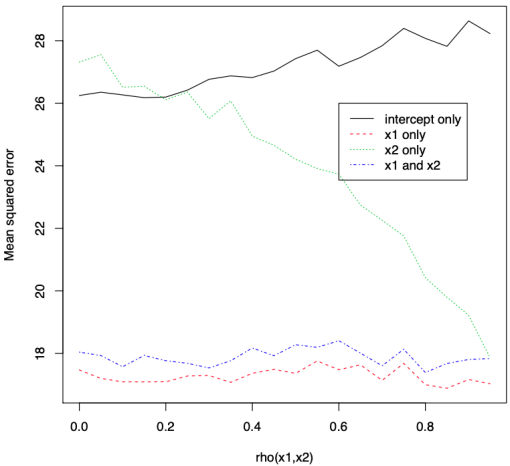
  where ϵ is from the $N(0, 16)$ distribution.
- We used a training sample size of $ntrain = 30$ and a test sample size of $ntest = 100,000$.

Example of biais-variance tradeoff (cont')

- We fit the following 4 models with the training data set :
  1) with no variables (intercept only)
  2) with X_1 only
  3) with X_2 only
  4) with X_1 and X_2.
- Each model was fit and the mean squared error (MSE) was evaluated with the test set.
- The whole process was repeated 20 times and the results averaged.

Example of biais-variance tradeoff (cont')

- The following figure presents the MSE, as a function of ρ, for the 4 models.



- We see that the model with X_1 only (red curve) is always the best one - better than the true model with both variables (blue line).

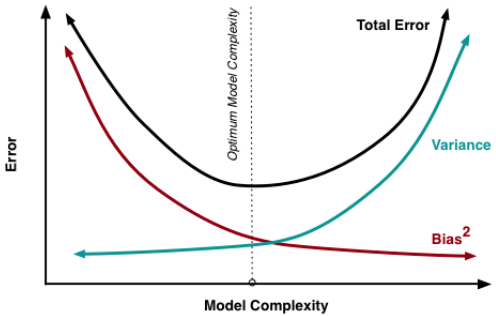
Example of biais-variance tradeoff (cont')

- The predictive signal coming from X_2 ($\beta_2 = 0.4$) is a lot smaller than X_1 ($\beta_1 = 3$).
- The cost, in terms of increased variance, of estimating the parameter of X_2 is larger than the predictive power it provides. Hence it is better not to use it.
- For small values of ρ, the intercept only model is even better than the model with X_2 only.
- When ρ increases, X_1 and X_2 become more and more alike and it becomes more difficult to divide their total effect between them. Near the end, for high values of ρ, they are almost the same variable. At this stage, using X_2 alone or both variables provide the same performance.

Example of biais-variance tradeoff: lessons to learn

- This example illustrates a few things.
- The true model is not always the best one in terms of predictive performance.
- The predictive value of a variable depends not only on its individual effect (its parameter here), but also on its correlation with other predictors.

Illustration of the biais-variance tradeoff



Source: http://scott.fortmann-roe.com/docs/BiasVariance.html

Agenda

Introduction

1. Introduction
2. Basic concepts: bias-variance tradeoff
3. **Basic concepts: maximum likelihood estimation**
4. Model selection
   - Likelihood-based criteria (AIC / BIC)
   - Sample-splitting methods
   - Cross-validation
   - 1 SE rule with cross validation
5. Generalized Linear Models

Maximum likelihood estimation

- Maximum likelihood estimation is a technique used to estimate the parameter(s) of a model
- It's a parametric method, i.e. we suppose a distribution on the data

Definition of the likelihood function

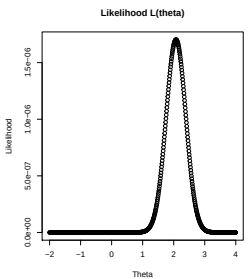
- Suppose that we observe a sample of data $Y_1, Y_2, \ldots, Y_n$
- Suppose that each observation Y_i follows a distribution that depends on a parameter θ. For example, let's say $Y_i \sim \mathcal{N}(\theta, 1)$.
- The likelihood function, denoted $L(\theta)$, is a function of the θ parameter given the observed data Y.
- The values taken by $L(\theta)$ tell us how likely the parameter is equal to θ, given the observed data.

Example of a likelihood function

- Let's consider for example $n = 10$ and suppose that we observe $Y_1, Y_2, \ldots, Y_{10}$ as follows:

1.43	1.76	3.55	2.07	2.12	3.71	2.46	0.73	1.31	1.55

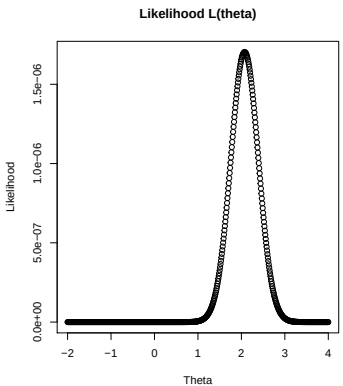
- If we assume that $Y_i \sim \mathcal{N}(\theta, 1)$ for all i, the likelihood function for the parameter θ is:



This figure tells us that it's much more likely that $\theta = 2$ than $\theta = 0$, for example.

Maximum likelihood estimation

- The main idea is to find the value of the parameter θ (for example) that maximizes the likelihood function.
- Given the observed data, the value $\hat{\theta}$ maximizing the likelihood is the most likely value of θ
- In our previous example, the maximum likelihood estimator is $\hat{\theta} = 2$.



Likelihood computation

- The likelihood function of a sample $Y_1, \ldots, Y_n$ is computed differently whether the variable is discrete or continue
- If the variable is discrete, the likelihood function is the probability function. Therefore:

$$\begin{aligned} L(\theta | Y_1 = y_1, \ldots, Y_n = y_n) &= P(Y_1 = y_1, \ldots, Y_n = y_n | \theta), \\ &= \prod_{i=1}^{n} P(Y_i = y_i | \theta), \end{aligned}$$

if the variables $Y_1, \ldots, Y_n$ are independent.

Likelihood computation (cont')

Probability functions for the common discrete distributions

Distribution	Notation	Parameter	Support	Probability function
Uniform	$Y \sim \mathcal{U}(a, b)$	(a, b)	$\{a, \ldots, b\}$	$P(Y = y) = \frac{1}{(b-a+1)}$
Bernouilli	$Y \sim \mathcal{B}(p)$	p	$\{0, 1\}$	$P(Y = 1) = p$ $P(Y = 0) = 1 - p$
Binomial	$Y \sim \mathcal{B}(n, p)$	(n, p)	$\{0, \ldots, n\}$	$P(Y = y) = \binom{n}{k} p^k (1-p)^{n-k}$
Poisson	$Y \sim \mathcal{P}(\lambda)$	λ	$[0, +\infty]$	$P(Y = y) = e^{-\lambda} \lambda^k / k!$

Likelihood computation (cont')

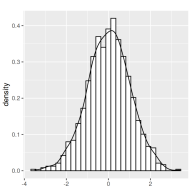
- If the variable is continue, the likelihood function is the density function, noted $f(y|\theta)$. Therefore we have:

$$L(\theta|Y_1 = y_1, \ldots, Y_n = y_n) \quad = \quad \prod_{i=1}^{n} f(y_i|\theta),$$

  if the variables $Y_1, \ldots, Y_n$ are independent.

- **The density is NOT the probability of Y.**

- A density function can be seen as the limit of an histogram when the number of classes is large



Likelihood computation (cont')

Density functions for the common continuous distribution

Distribution	Notation	Parameter	Support	Density function
Uniform	$Y \sim \mathcal{U}[a, b]$	(a, b)	$[a, b]$	$f(y) = \frac{1}{(b-a)}$
Exponential	$Y \sim \mathcal{E}(\lambda)$	λ	$[0, +\infty[$	$f(y) = \lambda e^{-\lambda y}$
Normal	$Y \sim \mathcal{N}(\mu, \sigma^2)$	(μ, σ)	$[-\infty, +\infty[$	$f(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(y-\mu)^2/2\sigma^2}$

Example 1: maximum likelihood

- Let's revisit our small example...
- Consider a sample of size $n = 10$ and suppose we observe $Y_1, Y_2, \ldots, Y_{10}$ as follows:

1.43	1.76	3.55	2.07	2.12	3.71	2.46	0.73	1.31	1.55

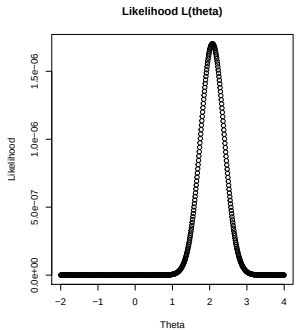
- If we suppose that $Y_i \sim \mathcal{N}(\theta, 1)$ for all i, then the likelihood of the parameter θ is:

$$L(\theta) = \prod_{i=1}^{n} f(y_i|\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-(y_i-\theta)^2/2\sigma^2},$$

where we replace σ by the value 1, and where we replace the values of y_i by $y_1 = 1.43, y_2 = 1.76, \ldots, y_{10} = 1.55$

Example 1: maximum likelihood (cont')

- The function $L(\theta)$ depends only of the parameter θ
- This function can be maximized and we find $\hat{\theta} = 2$.



- **Note**: it is often easier to maximize the log-likelihood. The solution can be computed numerically (approximation) or analytically (exact formula) in some simple cases, as it is the case in this example.

Example 2: maximum likelihood

- Let's look at an example of a basic linear regression.
- Consider a target variable Y and p predictors $X_1, X_2, \ldots, X_p$.
- We assume the model:

$$\begin{aligned} Y|X_1, X_2, \ldots, X_p & \quad \text{is normally distributed.} \\ E[Y|X_1, X_2, \ldots, X_p] = & \quad \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p \\ Var[Y|X_1, X_2, \ldots, X_p] = & \quad \sigma^2 \end{aligned}$$

- We have independent observations $(y_i, x_{1i}, \ldots x_{pi})$ for $i = 1, \ldots, n$.
- In this case $\theta = (\beta_0, \beta_1, \ldots, \beta_p, \sigma^2)$.

Example 2: maximum likelihood (cont')

- In our example, the density of an observation Y_i is

$$f(y_i|x_1, x_2, \ldots, x_p) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[\frac{-1}{2\sigma^2}\left(y_i - (\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p)\right)^2\right].$$

- Hence, for our n independent data, the log-likelihood function is

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} f(y_i|x_1, x_2, \ldots, x_p) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[\frac{-1}{2\sigma^2}\left(y_i - (\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p)\right)^2\right]. \end{aligned}$$

Exemple 2: maximum de vraisemblance (suite)

- As we said earlier, it's easier to maximize the log-likelihood given as:

$$l(\theta) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi}))^2.$$

- Maximizing this function with respect to $\beta_0, \ldots, \beta_p$ is equivalent to minimizing

$$\sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi}))^2.$$

- This is equivalent to the least-squares (LS) criterion !
- Hence, the maximum likelihood estimators (MLE) of the regression parameters in a basic linear regression model under normality are the same as the LS estimators.

Example 2: maximum likelihood (cont')

- Moreover, the MLE for σ^2 is

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

where $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \ldots \hat{\beta}_p x_{pi}$ are the fitted values from the model and the $\hat{\beta}$ are the MLE of the regression parameters.

Properties of the MLE

- Suppose we have n independent observations and a model for them, involving a parameter θ.
- Under certain conditions, the MLE of θ, call it $\hat{\theta}_{mle}$ has the following properties:
  1. $\hat{\theta}_{mle}$ is convergent, that is, $\hat{\theta}_{mle} \to \theta$, as $n \to \infty$.
  2. $\hat{\theta}_{mle}$ is asymptotically normally distributed.
  3. $\hat{\theta}_{mle}$ has the smallest variance among all estimators that are convergent and asymptotically normally distributed.

Agenda

Introduction

1. Introduction
2. Basic concepts: bias-variance tradeoff
3. Basic concepts: maximum likelihood estimation
4. **Model selection**
   - **Likelihood-based criteria (AIC / BIC)**
   - Sample-splitting methods
   - Cross-validation
   - 1 SE rule with cross validation
5. Generalized Linear Models

Variable selection

- It is important to keep in mind that how to approach the variable selection problem depends on the goal of the analysis.
  - Are we interested in having the best predictive model, or to find the true subset of variables related to the target?
  - We will come back to this aspect ...
- When the goal is to build a good prediction model, and many competing models are available, it is important to have criteria and/or techniques to help us select a model.
- In this section, we provide a brief review of some of them:
  - Likelihood-based criteria (AIC / BIC)
  - Sample-splitting methods
  - Cross-validation
  - 1 SE rule with cross validation

Likelihood-based criteria for model selection

- Two well-known model selection criteria based on likelihood:
  - The AIC (Akaike Information Criterion)
  - the BIC (Bayesian Information Criterion)
- They are defined by

$$\begin{aligned} \text{AIC} &= -2LL(\hat{\theta}_{mle}) + 2p, \\ \text{BIC} &= -2LL(\hat{\theta}_{mle}) + \log(n)p. \end{aligned}$$

where:

- $LL(\theta) = \log(L(\theta))$ is the log-likelihood function
- $\hat{\theta}_{mle}$ is the MLE
- p is the number of parameters that were estimated (i.e., the number of elements in θ)
- n is the sample size

Likelihood-based criteria for model selection (cont')

$$\begin{aligned} \text{AIC} &= -2LL(\hat{\theta}_{mle}) + 2p, \\ \text{BIC} &= -2LL(\hat{\theta}_{mle}) + \log(n)p. \end{aligned}$$

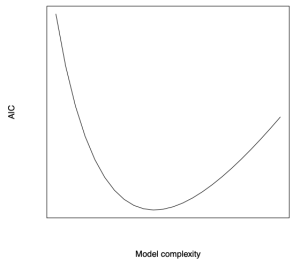
- The value $-2LL(\hat{\theta}_{mle})$ is a measure of the quality of the fit of the model.
- The quality of the fit can typically be improved by augmenting the number of parameters.
- Hence, in general, we can typically reduce $-2LL(\hat{\theta}_{mle})$ by adding more parameters.
- This is why $-2LL(\hat{\theta}_{mle})$ cannot be used alone as a model selection method.
- The terms $2p$ and $\log(n)p$ are penalty terms that try to avoid overfitting.

AIC versus BIC

- Smaller values of these criteria indicate better models.
- AIC and BIC can be seen as criteria that try to find a good compromise between the model fit and the number of parameters required to achieve the fit.
- The penalty of the BIC is greater than the one of the AIC (at least as soon as $n > 7$ which will be always the case).
- This means that the BIC will always select a model with a complexity less or equal than the one selected by the AIC
- This means that the number of parameters for the best model according to BIC will be less or equal to the number of parameters of the best model according to AIC.

Behavior of the AIC/BIC

- Typically, the behavior of the AIC (or BIC), as a function of the model complexity, will be as in the next figure.



- A too simplistic model is not good and a lot of gain are made early by increasing model complexity.
- But at some point, adding more complexity will begin to overfit the data.
- But the rate at which the model deteriorates after we begin to overfit is less than the rate at which it improves when we add complexity to an underfit model.

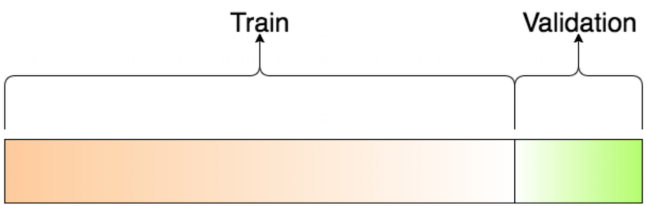
Agenda

Introduction

1. Introduction
2. Basic concepts: bias-variance tradeoff
3. Basic concepts: maximum likelihood estimation
4. **Model selection**
   - Likelihood-based criteria (AIC / BIC)
   - **Sample-splitting methods**
   - Cross-validation
   - 1 SE rule with cross validation
5. Generalized Linear Models

Sample-splitting methods

- Criteria like AIC and BIC are related to the MLE method.
- A predictive model (learner) is not always estimated via MLE.
- More general model selection methods are required.
- A very simple and universal way to compare the performance of competing models is to split the data in two parts: 1) training data, 2) validation data.
- The models are fit using the training data, then the performance of each of them is estimated using the validation data.



Sample-splitting methods to compare models

- With a continuous Y, the mean-squared error (MSE) could be used to compare models
  - The model with the smallest MSE on the validation sample would be declared the best one.
- For a categorical response, the misclassification rate on the validation sample could be the criterion.
  - The model with the smallest misclassification rate on the validation sample would be the best one.
- More generally, a general loss matrix that gives different values to different types of error could also be used.
- Note that it is not required to use a penalized criterion (like the AIC) because the validation set acts as new data.

Sample-splitting methods to estimate error

- If it is also required to get an estimate of the error, a better strategy is to split the data in three parts: 1) training data, 2) validation data, 3) test data.
- The first two parts are used as described before.
- The third part (test data) is only used to estimate the true error of the final selected model
- The test data is necessary to get a valid estimation of the error of the selected model
- When the validation data is used to evaluate many models, the one selected will tend to overfit the validation data.

Sample-splitting methods: summary

- Selecting the model with the validation data is a good strategy.
- But estimating the true error of the selected model with the validation data will usually underestimate the error.
- The test data should be used to estimate the true error of the selected model.
- In the end, when a model is selected, it should be fit again using all the observations available. This model becomes the one to use to predict new data.

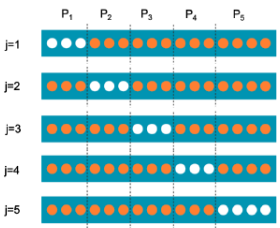
Agenda

Introduction

1. Introduction
2. Basic concepts: bias-variance tradeoff
3. Basic concepts: maximum likelihood estimation
4. **Model selection**
   - Likelihood-based criteria (AIC / BIC)
   - Sample-splitting methods
   - **Cross-validation**
   - 1 SE rule with cross validation
5. Generalized Linear Models

Cross-validation

- Data splitting (in 2 or 3 parts) is easy to implement but a fairly large data set is required.
- Cross-validation reproduces the same idea in a more clever way, but is more computer intensive.
- The basic idea is to partition the training data into subsets, usually 5 or 10 (5-fold or 10-fold cross-validation):
  - remove one subset at a time
  - estimate the model with the remaining subsets pooled together
  - compute the predictions (or any other required quantity) on the subset that was not used to fit the model.
  - Repeat this for all subsets and collect the results.



Cross-validation

- Note that each observation gets predicted exactly once, in the iteration where it was part of the removed subset.
- Hence, in the end, we have a single prediction for each observation. We can then compute the performance criterion, e.g. mean-squared error, or misclassification rate, using these out-of sample predictions.
- To reduce the variance related to the specific partition of the data, we can repeat the cross-validation process many times and average the results. For example repeat 10-fold cross-validation 20 times.

Agenda

Introduction

1. Introduction
2. Basic concepts: bias-variance tradeoff
3. Basic concepts: maximum likelihood estimation
4. **Model selection**
   - Likelihood-based criteria (AIC / BIC)
   - Sample-splitting methods
   - Cross-validation
   - **1 SE rule with cross validation**
5. Generalized Linear Models

Cross-validation

- The obvious way to select a model with cross-validation is to select the one which produces the smallest value of the estimated error.
- Another popular way is to select a simpler model whose estimated error is not too far away from the minimum one.
- The 1-SE rule (one standard error rule) is one such way.

The 1-SE rule

- Assume we have estimated, by cross-validation, the error of many models indexed by a tuning parameter θ which represents the complexity of the model.
- Larger values of θ indicate simpler models.
- For example, θ could be:
  - the tuning parameter λ in the lasso,
  - the value $n - k$ where k is the number of variables in the model (hence when k increases, then θ decreases and the model's complexity increases).
- Assume we have estimated the error by cross-validation at M different values of θ, which are $\theta_1, \theta_2, \ldots, \theta_M$.
- Call these estimations $CV(\theta_1), CV(\theta_2), \ldots, CV(\theta_M)$.
- Assume we also have the standard errors of these estimations, $SE(\theta_1), SE(\theta_2), \ldots, SE(\theta_M)$

The 1-SE rule (cont't)

- As mentioned above, the obvious way is to select the model with minimum estimated error, that is, the model for which θ is

$$\hat{\theta}_{min} = \underset{\theta \in \{\theta_1, \theta_2, \ldots, \theta_M\}}{\arg\min} \; CV(\theta).$$

- The 1-SE rule selects the model with the largest value of θ such that

$$CV(\theta) \leq CV(\hat{\theta}_{min}) + SE(\hat{\theta}_{min}).$$

- That is, the simplest model with an error within 1 standard error of the error of the model with the smallest error.

- Note that this is an ad-hoc rule that has no theoretical justification. But it is often mentioned and sometimes used in practice, especially with tree-based methods.

The 1-SE rule (cont't)

- The standard errors themselves can be obtained the following way.
- Assume we perform K-fold cross-validation.
- For a given θ, let $CV_k(\theta)$ be the estimated error for the k^{th} fold, $k = 1, 2, \ldots, K$. For example, this could be the mean squared error.
- The standard error of $CV(\theta)$ is then

$$SE(\theta) = \frac{\sqrt{Var(CV_1(\theta), CV_2(\theta), \ldots, CV_K(\theta))}}{\sqrt{K}} = \sqrt{\frac{\sum_{k=1}^{K}(CV(\theta_k) - \overline{CV})^2}{K(K-1)}},$$

  where

$$\overline{CV} = \frac{\sum_{k=1}^{K} CV(\theta_k)}{K}$$

  is the average error across the folds.

- Hence, it is based on the variability between the folds.

Toy example

- Suppose we are in the context of a regression and we want to compare 4 models:
  - Model 1: 1 variable
  - Model 2: 2 variables
  - Model 3: 3 variables
  - Model 4: 4 variables
- We want to select the best model
- **For each model**, we perform a 10-fold cross validation as follows:
  - For each fold k, we compute the MSE: MSE^1,...,MSE^{10}
  - Using these 10 MSE's, one can compute the standard deviation of the MSE and the average MSE of the model.

Toy example (cont')

- Suppose we obtain the following results:

Model	Mean MSE	MSE standard deviation
1	45	8
2	40	5
3	35	7
4	38	6

- The best model according to the MSE criterion is model 3, with a MSE=35 and standard deviation 7
- With the 1-SE rule, the best model is the one with the smallest number of variables such that MSE$\leq 35 + 7 = 42$
- The winner is model 2.

Agenda

Introduction

1. Introduction
2. Basic concepts: bias-variance tradeoff
3. Basic concepts: maximum likelihood estimation
4. Model selection
   - Likelihood-based criteria (AIC / BIC)
   - Sample-splitting methods
   - Cross-validation
   - 1 SE rule with cross validation
5. **Generalized Linear Models**

Generalized Linear Models (GLMs)

- An important class of parametric models that encompasses linear, Poisson, and logistic regression is the class of generalized linear models (GLMs).
- As usual, we denote by Y the response and by $\mathbf{X} = (X_1, X_2, \ldots, X_p)$ the associated vector of covariates.
- The goal is to model the mean of Y as a function of $\mathbf{X}$. This is noted $\mu = E[Y|\mathbf{X}]$.

Generalized Linear Models (GLMs) (cont')

- Two components are required to define a GLM:
  1. An assumed probability distribution for Y.
  2. A continuous function g linking the conditional mean to the covariates via:

$$g(\mu) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$$

The function g is called the "link" function.

Examples of GLMs

- To fix ideas, let's look at three particular cases:
  - The linear regression under normality
  - The logistic regression
  - The Poisson regression

Examples of GLMs: linear regression

- Recall the GLM definition:

$$g(\mu) \quad = \quad \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$$

- Assume now that Y is from the normal distribution
- Take $g(\mu) = \mu$ (identity function) as the link function
- Then the corresponding GLM specifies that

$$g(\mu) = \mu = E[Y|\mathbf{X}] = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p,$$

- We see that this GLM is exactly the same as an ordinary linear regression model, that is we assume that

$$Y \sim \mathcal{N}(\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p, \sigma^2)$$

Examples of GLMs: logistic regression

- Assume now Y is binary and can take values 0 and 1.
- For such a variable, it is straightforward to see that

$$E[Y|\mathbf{X}] = P(Y = 1|\mathbf{X}),$$

- Hence, specifying a model for $E[Y|\mathbf{X}]$ amounts to specify one for $P(Y = 1|\mathbf{X})$.

Examples of GLMs: logistic regression (cont')

- Define the logit link function to be

$$g(\mu) = \log\left(\frac{\mu}{1-\mu}\right).$$

- If we use this function as the link function in the GLM, then we have

$$\log\left(\frac{P(Y=1|\mathbf{X})}{1-P(Y=1|\mathbf{X})}\right) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p.$$

- We can express this last equation as a function of $P(Y=1|\mathbf{X})$ by

$$P(Y=1|\mathbf{X}) = \frac{1}{1+\exp(-(\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p))}.$$

- This is a logistic regression model.

Examples of GLMs: Poisson regression

- Assume now Y can take the values $0, 1, 2, \ldots$ and is from the Poisson distribution.
- The Poisson distribution has a single parameter, call it $\lambda > 0$, and its probability distribution is given by

$$P(X = x) = \frac{\lambda^x \exp(-\lambda)}{x!}, \quad \text{for} \quad x = 0, 1, \ldots$$

- The mean of the Poisson distribution is:

$$E[X] = \lambda.$$

- If we take the log link function $g(\mu) = \log(\mu)$, the corresponding GLM assumes that Y follows a Poisson distribution with

$$\log(E[Y|\mathbf{X}]) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p,$$

or equivalently,

$$E[Y|\mathbf{X}] = \exp(\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p).$$