Classification

Logistic Model

Ana Karina Fermin

Université Paris Nanterre

ISEFAR
http://fermin.perso.math.cnrs.fr/

1. k Nearest-Neighbors ✓
2. K-Fold cross validation, Model selection ✓
3. Generative Modeling (Naive Bayes, LDA, QDA) ✓
4. Logistic Modeling
5. SVM
6. Tree Based Methods (M. Zetlaoui course, Apprentissage)

- Direct modeling of $Y|x$.

The Binary logistic model ($Y \in \{-1, 1\}$)

$$p_{+1}(\mathbf{x}) = \frac{e^{\beta^t \varphi(\mathbf{x})}}{1 + e^{\beta^t \varphi(\mathbf{x})}}$$

where $\varphi(x)$ is a transformation of the individual $\mathbf{x}$

- In this model, one verifies that
$$p_{+1}(\mathbf{x}) \geq p_{-1}(\mathbf{x}) \quad \Leftrightarrow \quad \beta^t \varphi(\mathbf{x}) \geq 0$$
- True $Y|x$ may not belong to this model $\Rightarrow$ maximum likelihood of β only finds a good approximation!
- Binary Logistic classifier:
$$\widehat{f}_L(\mathbf{x}) = \begin{cases} +1 & \text{if } \widehat{\beta}^t \varphi(\mathbf{x}) \geq 0 \\ -1 & \text{otherwise} \end{cases}$$

where $\widehat{\beta}$ is estimated by maximum likelihood.

- Logistic model: approximation of $\mathcal{B}(p_1(\mathbf{x}))$ by $\mathcal{B}(h(\beta^t\varphi(\mathbf{x})))$ with $h(t) = \frac{e^t}{1+e^t}$.

Opposite of the log-likelihood formula

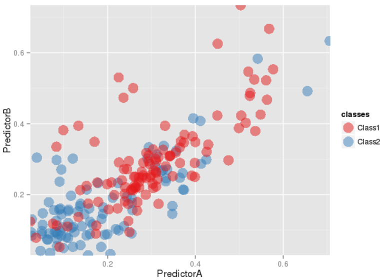
$$
\begin{aligned}
&-\frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{1}_{y_i=1}\log(h(\beta^t\varphi(\mathbf{x}))) + \mathbf{1}_{y_i=-1}\log(1-h(\beta^t\varphi(\mathbf{x})))\right)\\
&= -\frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{1}_{y_i=1}\log\frac{e^{\beta^t\varphi(\mathbf{x})}}{1+e^{\beta^t\varphi(\mathbf{x})}} + \mathbf{1}_{y_i=-1}\log\frac{1}{1+e^{\beta^t\varphi(\mathbf{x})}}\right)\\
&= \frac{1}{n}\sum_{i=1}^{n}\log\left(1+e^{-y_i(\beta^t\varphi(\mathbf{x}))}\right)
\end{aligned}
$$

- Convex function in β!
- **Remark:** You can also use your favorite parametric model instead of the logistic one...

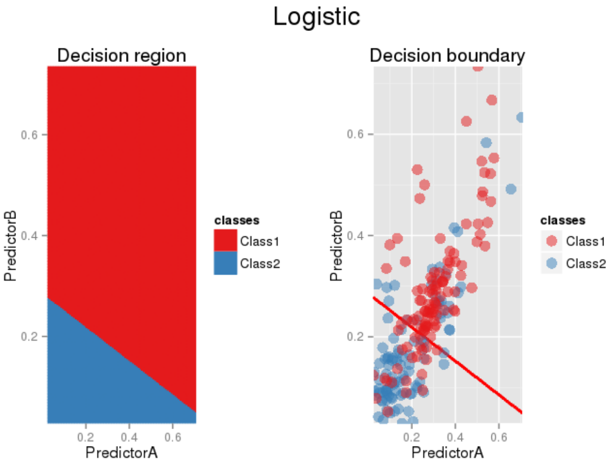
Synthetic Dataset

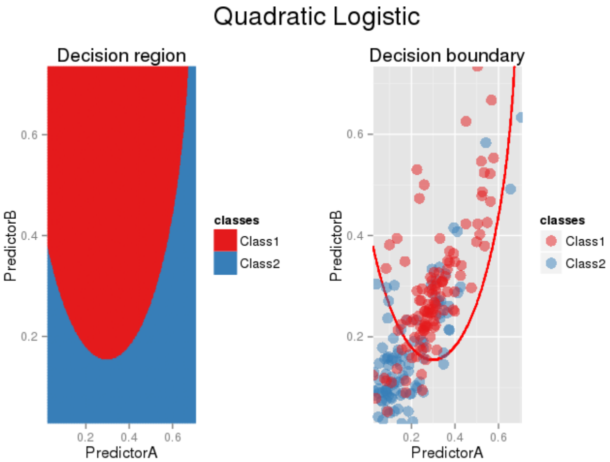
- Two features/covariates.
- Two classes.

- Dataset from *Applied Predictive Modeling*, M. Kuhn and K. Johnson, Springer
- Numerical experiments with **R**.



Example: Logistic





Quadratic Logistic
Decision region
Decision boundary
PredictorA
PredictorB
0.2
0.4
0.6
classes
Class1
Class2

1. Model Selection
2. Prediction Errors

1. Given Y a variable to explain by d variables $X_1, \ldots, X_d$, how to select (systematically) the most interesting subset of variables to do the prediction?

Variable selection

Find automatically a sub-group of variables to explain Y.

2. More generaly, given k models $\mathcal{M}_1, \ldots, \mathcal{M}_k$, which one to use?

Model selection

Criterion to compare the performance of different models.

- Assume we have two competing models $\mathcal{M}_p$ (with p parameters) and $\mathcal{M}_q$ (with q parameters) such that $\mathcal{M}_p \subset \mathcal{M}_q$.
- Can we test if $\mathcal{M}_p$ is sufficient?

Example (p=2 and q=4)

- Models:
  - $\mathcal{M}_p$: $\mathrm{logit}\, p_\beta^{(p)}(x) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$
  - $\mathcal{M}_q$: $\mathrm{logit}\, p_\beta^{(q)}(x) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$
- Test

$$H_0 : \beta_3 = \beta_4 = 0 \quad \text{against} \quad H_1 : \beta_3 \neq 0 \quad \text{or} \quad \beta_4 \neq 0$$

Deviance (`residual deviance` sous R)

- Log-likelihood $\mathcal{M}_p$: $\mathcal{L}_p = \log(L_p(\hat{\beta}, \mathcal{D}_n))$
- Log-likelihood of model $\mathcal{M}_q$: $\mathcal{L}_q = \log(L_q(\hat{\beta}, \mathcal{D}_n))$
- Deviance between the two models:

$$\mathcal{D}_{q-p} = 2(\mathcal{L}_q - \mathcal{L}_p) = 2(\log(L_q(\hat{\beta}, \mathcal{D}_n)) - \log(L_p(\hat{\beta}, \mathcal{D}_n)))$$

Asymptotically under H_0 : $(\mathcal{D}_{q-p}) \sim \chi^2(q-p)$

Under R : If `W` and `V` are two objects obtained with `glm` such that `W` is a submodel of `V`, the command

```
anova(W,V,test="Chisq")
```

performs this test.

- Let $\mathcal{M}$ be a generic logistic model and denote p its number of parameters.
- Let $\hat{\beta}$ be the ML estimate in this model $\mathcal{M}$.

- The AIC and BIC consist in minimizing

$$-2 \times \log(L(\hat{\beta}, \mathcal{D}_n)) + \kappa(n) \times p$$

  over all models.
- Different choices for the factor $\kappa(n)$:
  - AIC : $\kappa(n) = 2$.
  - BIC : $\kappa(n) = \log n$.
- The BIC criterion leads to the selection of a model with a smaller dimension than AIC.

Prediction for a new individual:

- A new individual x_{new} appears and we want to predict if he has the disease ($y_{\text{new}} = 1$) or not ($y_{\text{new}} = 0$).
- We have estimated the logistic coefficients $\widehat{\beta}$ so that

$$\mathbb{P}\left\{Y = 1 | X\right\} \approx \frac{e^{\widehat{\beta}^T \mathbf{X}}}{1 + e^{\widehat{\beta}^T \mathbf{X}}}. \tag{1}$$

- Any ideas to predict y_{new}?

Prediction (threshold = 0.5)

Then,

- If $\widehat{\mathbb{P}}(y = \mathrm{Yes}|X) > 0.5$, we predict $y = \mathrm{Yes}$;
- If $\widehat{\mathbb{P}}(y = \mathrm{Yes}|X) \leq 0.5$, we predict $y = \mathrm{No}$.

Confusion Matrix : Cross table of the prediction vs the truth

```
##      pred
## CHD   No Yes
##   No  45  12
##   Yes 14  29
```

Prediction error : $(14 + 12)/100 = 0.26$

Prediction Errors

Classical Metrics

Actual Class \ Predicted Class	Yes	No
Yes	TP	FN
No	FP	TN

Score

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

Scores

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

$$\left.\begin{array}{c}\text{True Positive Rate}\\ \text{Sensitivity}\\ \text{Recall}\end{array}\right\} = \frac{\text{TP}}{\#(\text{real P})} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\left.\begin{array}{c}\text{False Positive Rate}\\ 1 - \text{Specificity}\end{array}\right\} = \frac{\text{FP}}{\#(\text{real N})} = \frac{\text{FP}}{\text{FP} + \text{TN}}$$

$$\text{Precision} = \frac{\text{TP}}{\#(\text{predicted P})} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{F-Score} = 2\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

To label digits:

- True label $y_i \in \{0, \ldots, 9\}$,
- Predicted label $\widehat{y}_i \in \{0, \ldots, 9\}$,
- Confusion matrix is thus of size 10×10.

Confusion matrix