STATS 415: Classification – Logistic regression

Prof. Liza Levina

Department of Statistics, University of Michigan

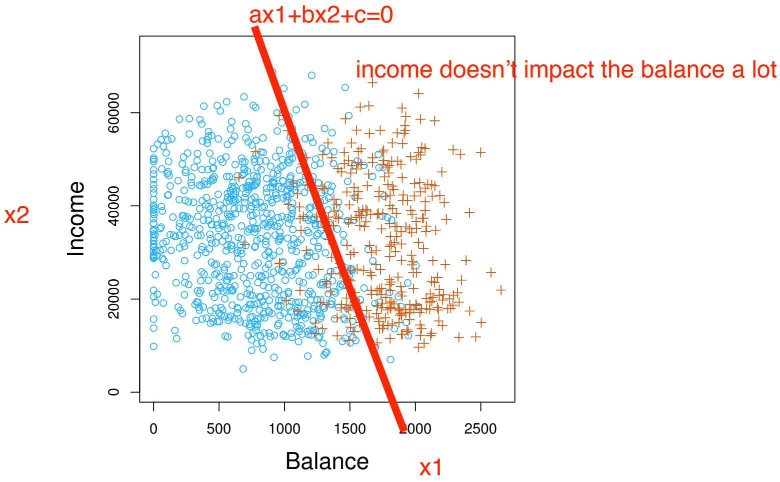
Recall the general principles

- The optimal classifier (Bayes Rule): picks c_k such that $P(Y = c_k|X = x) \propto p_k(x)\pi_k$ is maximized. Requires knowing true π_k and $p_k(x)$.
- Generative classification methods: estimate π_k, $p_k(x)$ from training data (parametrically or non-parametrically), then plug in into the Bayes rule
- Discriminative classification methods: estimate $P(Y = k|X = x)$ directly, without going through the Bayes rule

Logistic regression

- A discriminative parametric method
- Two-class case: $K = 2$, $y \in \{c_1, c_2\}$
- Example: Credit card default data
  - Possible prediction variables are: annual income, monthly credit card balance, mortgage payments, etc
  - The response variable (Default) is categorical: Yes or No
  - We would like to predict which customers are likely to default
  - We would also like to learn about the relationship between y and x

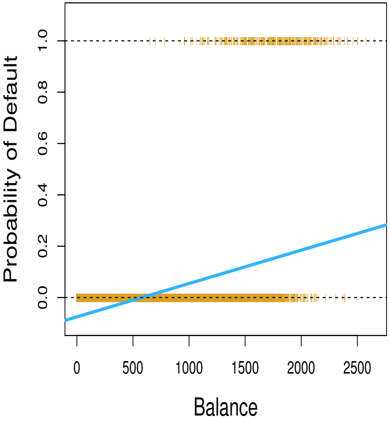
Credit card default data



Why not linear regression?

- Code the values of y using 0 and 1. Regress y on x?
- The regression function $\beta_0 + \beta^\top x$ can take on any value between negative and positive infinity.
- In a classification problem, y can only take on two possible values: 0 or 1. logistic regression (why not linear regression?)
- The point "cloud" has two y values only; the line is not a good model even in the range of x where y remains between 0 and 1

Linear regression vs logistics regression



The Logistic function

- Instead of trying to predict y itself, let's model $P(Y = c_k | X = x)$
- Two-class problem – just need to compare two values, $P(Y = c_1 | X = x)$ and $P(Y = c_2 | X = x)$
- Will often write $P(Y = c_k | X = x) = P(Y = c_k | x) = P(c_k | x)$
- Use the logit transformation (logistic function):

$$\log \frac{P(Y = c_1 | x)}{P(Y = c_2 | x)} = \beta_0 + \beta^\mathsf{T} x$$

- Can solve for probabilities:

$$\begin{aligned} P(Y = c_1 | x) &= \frac{e^{\beta_0 + \beta^\mathsf{T} x}}{1 + e^{\beta_0 + \beta^\mathsf{T} x}} \\ P(Y = c_2 | x) &= \frac{1}{1 + e^{\beta_0 + \beta^\mathsf{T} x}} \end{aligned}$$

- The probabilities are automatically between 0 and 1, and $P(Y = c_1 | x) + P(Y = c_2 | x) = 1$.

Writing the likelihood of binary variables

- Would like to fit the model by maximum likelihood estimation
- Let $Y = 1$ with probability p or 0 with probability $1 - p$

$$P(Y = y) = \begin{cases} p, & \text{if } y = 1 \\ 1 - p, & \text{if } y = 0 \end{cases} \quad = \quad p^y (1-p)^{1-y}$$

- Observe an i.i.d. sample from this distribution: $y_1, y_2, \ldots, y_n$
- Likehood as a function of p:

$$L(p; y_1, \ldots, y_n) = \prod_{i=1}^{n} p^{y_i} (1-p)^{1-y_i}$$

- Log-likelihood:

$$\ell(p; y_1, \ldots, y_n) = \log L = \sum_{i=1}^{n} y_i \log p + (1 - y_i) \log(1 - p)$$

Fitting logistic regression

- Maximum likelihood estimation
- Write $\beta = (\beta_0, \beta)$
- Write x for $(1, x)$ (add the intercept column)
- Conditional log-likelihood of y given x

$$
\begin{aligned}
\ell(\beta) &= \sum_{i=1}^{n} \log P(Y = y_i | X = x_i; \beta) \\
&= \sum_{i=1}^{n} [y_i \log P(c_1 | x_i; \beta) + (1 - y_i) \log P(c_2 | x_i; \beta)] \\
&= \sum_{i=1}^{n} [y_i (\beta^\mathsf{T} x_i) - \log(1 + e^{\beta^\mathsf{T} x_i})]
\end{aligned}
$$

Maximizing the likelihood

- Unlike with linear regression, there is no closed form solution for $\hat{\beta}$
- Score equation: take derivative with respect to β and set to 0

$$\frac{\partial \ell(\beta)}{\partial \beta} = \sum_{i=1}^{n} x_i (y_i - p(x_i;\beta)) = 0$$

where $p(x_i;\beta) = e^{\beta^\mathsf{T} x_i}/(1+e^{\beta^\mathsf{T} x_i})$.

- There are $(p+1)$ equations, nonlinear in β – has to be solved numerically

The iteratively reweighted least squares algorithm

- Solve with Newton-Raphson algorithm, a standard iterative general numerical optimization algorithm
- For logistic regression, reduces to iteratively reweighted least squares, at each iteration solving a weighted least squares problem of the form

$$\beta^{\text{new}} = \arg\min_{\beta} (z - X\beta)^{\mathsf{T}} W (z - X\beta)$$

- Requires an initial value β^0; can use $\beta^0 = 0$
- Must be iterated until β does not change anymore; does not always converge.

Inference

- If the model is correct, $\hat{\beta}$ is consistent, that is, $\hat{\beta} \to \beta$ as the sample size n grows.
- The distribution of $\hat{\beta}$ converges to $N(\beta, (X^\top W X)^{-1})$.
- Thus $\hat{\beta}$ is asymptotically unbiased ($E\hat{\beta} \to \beta$) as $n \to \infty$.

Example: Credit Card Default Data

	Coefficient	Std Err	Z-value	p-value
Intercept	-10.87	0.4923	-22.08	$<$ 0.0001
balance	0.0057	0.0002	24.74	$<$ 0.0001
income	0.0030	0.0082	0.37	0.7115
student[Yes]	-0.6468	0.2362	-2.74	0.0062

Inference example

- Test the hypothesis $H_0 : \beta_{\text{balance}} = 0$.
- The p-value for balance is very small, so we can reject H_0 and conclude balance "helps" predict default
- The coefficient $\hat{\beta}_{\text{balance}}$ is positive, so we are "sure" that if the balance increases (with all other predictors being held constant), then the probability of default will increase.
- Prediction: A student with an average credit card balance of \$1,500 and an income of \$40,000 has an estimated probability of default

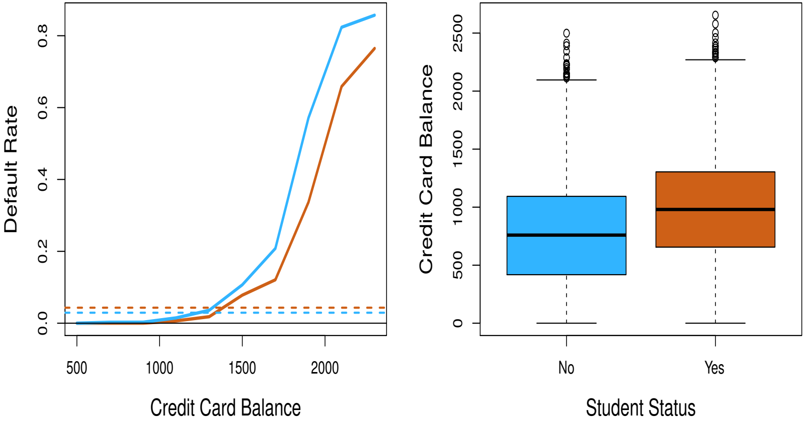
$$\widehat{P}(\text{default}|x) = \frac{e^{-10.87+0.0057\times1500+0.003\times40-0.6468}}{1+e^{-10.87+0.0057\times1500+0.003\times40-0.6468}} = 0.058$$

An Apparent Contradiction

	Coefficient	Std Err	Z-value	p-value
Intercept	-3.5041	0.0707	-49.55	< 0.0001
student[Yes]	0.4049	0.1150	3.52	0.0004

- A student is risker than a non-student if the credit card balance is not in the model.
- A student is less risky than a non-student with the same credit card balance, if the balance is in the model

Students vs Non-students



Multi-class (multinomial) logistic regression

- Use one class, say class K, as a reference

$$
\begin{aligned}
\log \frac{P(Y = c_1 | x)}{P(Y = c_K | x)} &= \beta_{10} + \beta_1^\mathsf{T} x \\
\log \frac{P(Y = c_2 | x)}{P(Y = c_K | x)} &= \beta_{20} + \beta_2^\mathsf{T} x \\
\vdots &= \vdots \\
\log \frac{P(Y = c_{K-1} | x)}{P(Y = c_K | x)} &= \beta_{(K-1)0} + \beta_{(K-1)}^\mathsf{T} x
\end{aligned}
$$

Logistic Regression vs LDA

- For LDA, can also calculate the logit of class odds for classes k and K:

the dimension of xT:1*p
the dimension of ∑:p*p
the dimension of mu: P*1

$$
\begin{aligned}
\log \frac{P(Y = c_k|x)}{P(Y = c_K|x)} &= x^\top \Sigma^{-1}(\mu_k - \mu_K) + \log \frac{\pi_k}{\pi_K} \\
&\quad - \frac{1}{2}(\mu_k + \mu_K)^\top \Sigma^{-1}(\mu_k - \mu_K) \\
&= \alpha_{k0} + \alpha_k^\top x
\end{aligned}
$$

=αk0+αk1x1+…αkpxp

- Logistic model:

$$
\log \frac{P(Y = c_k|x)}{P(Y = c_K|x)} = \beta_{k0} + \beta_k^\top x
$$

- Both are linear functions of x! Are they the same?

Where does this linearity come from

- LDA: linearity is a consequence of the Gaussian assumption for the class densities and the assumption of a common covariance matrix.
- For logistic regression, linearity is there by construction.
- The coefficients are estimated differently.

Common component: linear log-odds

- The joint density of (x, y) is

common component: linear log-odds

$$P(X = x, Y = c_k) = P(X = x)P(Y = c_k|X = x) = p(x)P(Y = c_k|X = x)$$

where $p(x)$ is the marginal density of the input x.

- For both LDA and logistic regression, the term $P(Y = c_k|x)$ has the same logit linear form

$$P(Y = c_k|x) = \frac{\exp(\beta_{k0} + \beta_k^\mathsf{T} x)}{1 + \sum_{k'=1}^{K} \exp(\beta_{k'0} + \beta_{k'}^\mathsf{T} x)}$$

Which model is more general?

LDA and logistic regression make different assumptions about $p(x)$.

- The logistic model leaves the marginal density of x arbitrary and unspecified.
- The LDA model assumes a Gaussian mixture density

$$p(x) = \sum_{k=1}^{K} \pi_k \cdot \phi(x; \mu_k, \Sigma)$$

- Logistic regression makes fewer assumptions about the data, and is more general.

Parameter estimation

Logistic regression

- Maximizing the conditional likelihood, the multinomial likelihood with probabilities $P(Y = c_k|x)$.
- The marginal density $p(x)$ is ignored (or rather estimated fully nonparametrically by the empirical distribution, which is a histogram with weight $1/n$ at each observation).

LDA

- Maximizing the full likelihood based on the joint density

$$P(x, Y = c_k) = \phi(x; \mu_k, \Sigma) \cdot \pi_k$$

- Marginal density does play a role.

Remarks

- LDA is easier to compute than logistic regression.
- If the true $f_k(x)$'s are Gaussian, LDA is better: logistic regression may lose up to 30% efficiency in error rate (Efron 1975).
- LDA uses all the data points to estimate the covariance matrix – more information but not robust against outliers.
- Logistic regression, through interatively reweighted least squares, down-weighs points far from the decision boundary; more robust.

Comparison of classification methods

- KNN
- Logistic regression
- LDA
- QDA

Logistic Regression vs LDA

- Main similarity: Both Logistic regression and LDA produce linear boundaries.
- Main difference: LDA assumes that the observations are drawn from the normal distribution with common variance in each class, while logistic regression does not.
- LDA is expected to outperform logistic regression if the normal assumption holds, otherwise logistic regression can outperform LDA.

KNN vs LDA / Logistic regression

- KNN takes a different approach: completely non-parametric
- No assumptions are made about the shape of the decision boundary
- Main advantage of KNN: deals well with non-linear and highly complex boundaries.
- Main disadvantage of KNN: no inference (no coefficients for the predictors or p-values). KNN vs LDA/Logistic
- We expect KNN to dominate both LDA and logistic regression when the decision boundary is highly non-linear.

QDA vs (LDA, Logistic Regression, and KNN)

- QDA is a compromise between the completely non-parametric KNN method and the linear LDA and logistic regression.
- The boundary is non-linear, but still of a specified form (quadratic)
- Also makes the normal assumption
- LIkely the best choice when the true decision boundary is:
    - Linear: LDA and logistic regression
    - Moderately non-linear: QDA
    - More complicated: KNN