

Penalized Logistic Regression

Stefanie Scheid

Max-Planck-Institute for Molecular Genetics

Computational Diagnostics

October 7, 2002

Literature

Eilers, PH, Boer, JM, Van Ommen, GJB, Van Houwelingen, HC (2001):
Classification of Microarray Data with Penalized Logistic Regression,
Proceedings of SPIE volume 4266: progress in biomedical optics and imaging,
2, 187-198.



Given data

| Y | X | | | |
|-----|-------|-------|-------|-----|
| 0 | 1.78 | -0.62 | 3.79 | ... |
| 0 | 0.39 | 1.73 | 2.09 | ... |
| 0 | -2.17 | 0.68 | 0.43 | ... |
| 1 | 1.89 | 3.62 | 2.90 | ... |
| 1 | 2.01 | 2.51 | 1.45 | ... |
| 1 | 3.81 | 1.90 | -0.82 | ... |



Linear vs. logistic regression

- Task: Find a model for

$$p = \text{Prob}(Y = 1 | X = x)$$

- Linear model:

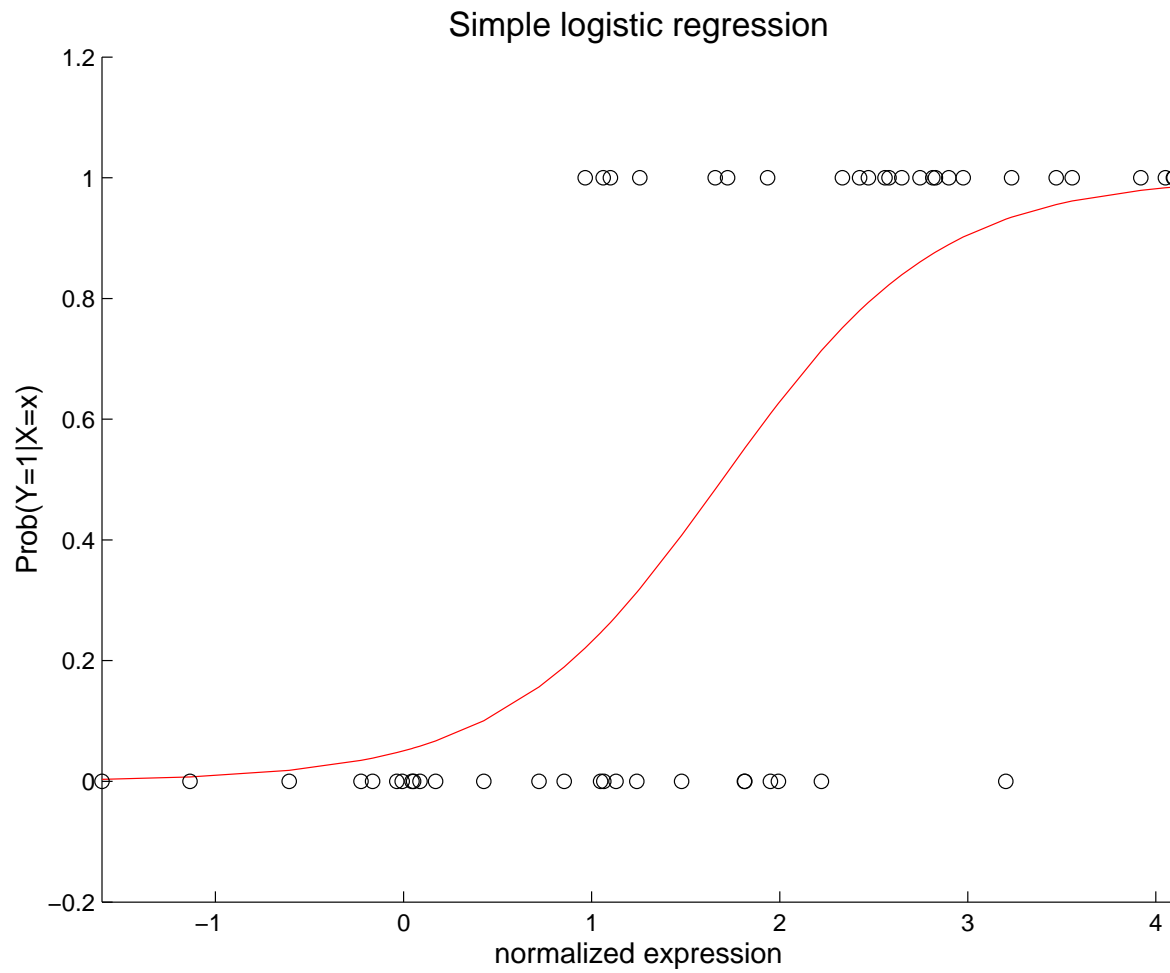
$$Y \in \{0, 1\} \quad \rightarrow \quad p = Y = \alpha + X \beta$$

- Logistic model:

$$\eta = \log \frac{p}{1-p} = \alpha + X \beta \quad \rightarrow \quad p = \frac{1}{1 + \exp(-\eta)} \in [0, 1]$$



Example



Adding a penalty term

- Likelihood:

$$L = \prod f(y, p) = \prod_{m} p^y (1 - p)^{1-y}$$

- Log-Likelihood:

$$\text{Log}L = \sum_{m} y \log p + \sum_{m} (1 - y) \log(1 - p)$$

- Penalized Log-Likelihood:

$$\text{Log}L^* = \text{Log}L - \lambda \sum_{n} \beta^2 / 2 \quad \rightarrow \quad \text{max!}$$

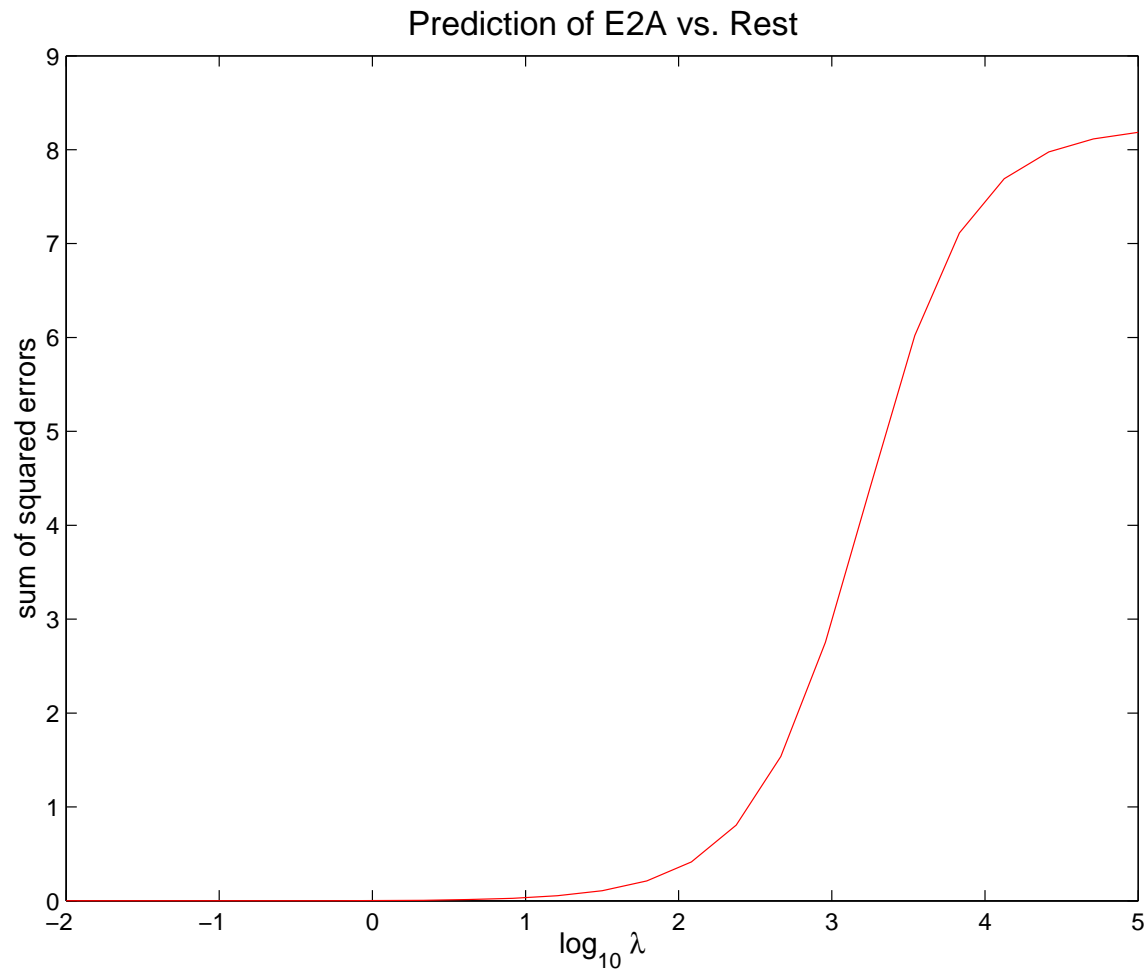


How to choose λ ?

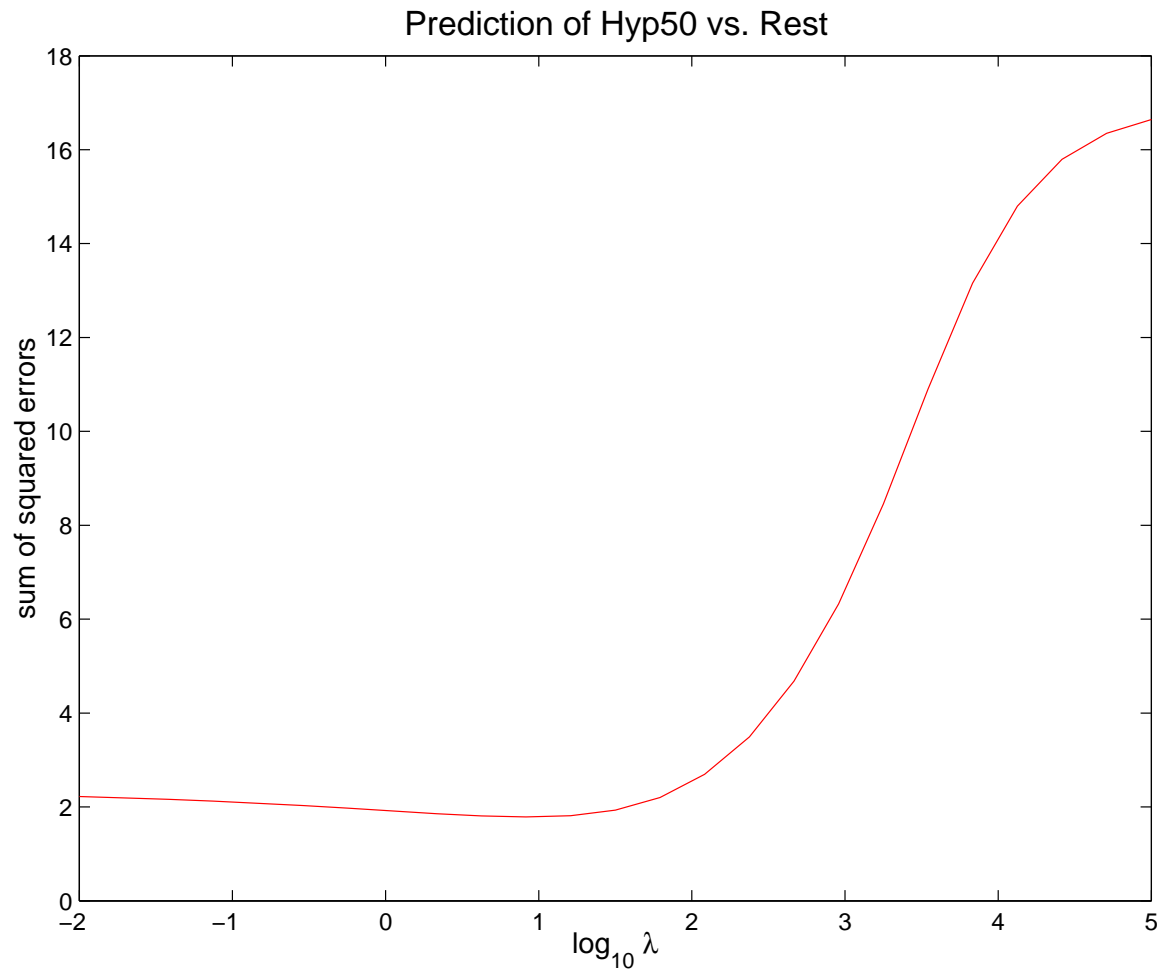
- Take λ that allows best prediction.
- Application: St. Jude data
 - training and test samples
 - preselection of 1000 genes (t-test)
 - $\lambda \in [10^{-2}, 10^5]$, take 25 linearly spaced values of $\log_{10} \lambda$



St. Jude data: E2A vs. Rest



St. Jude data: Hyp50 vs. Rest

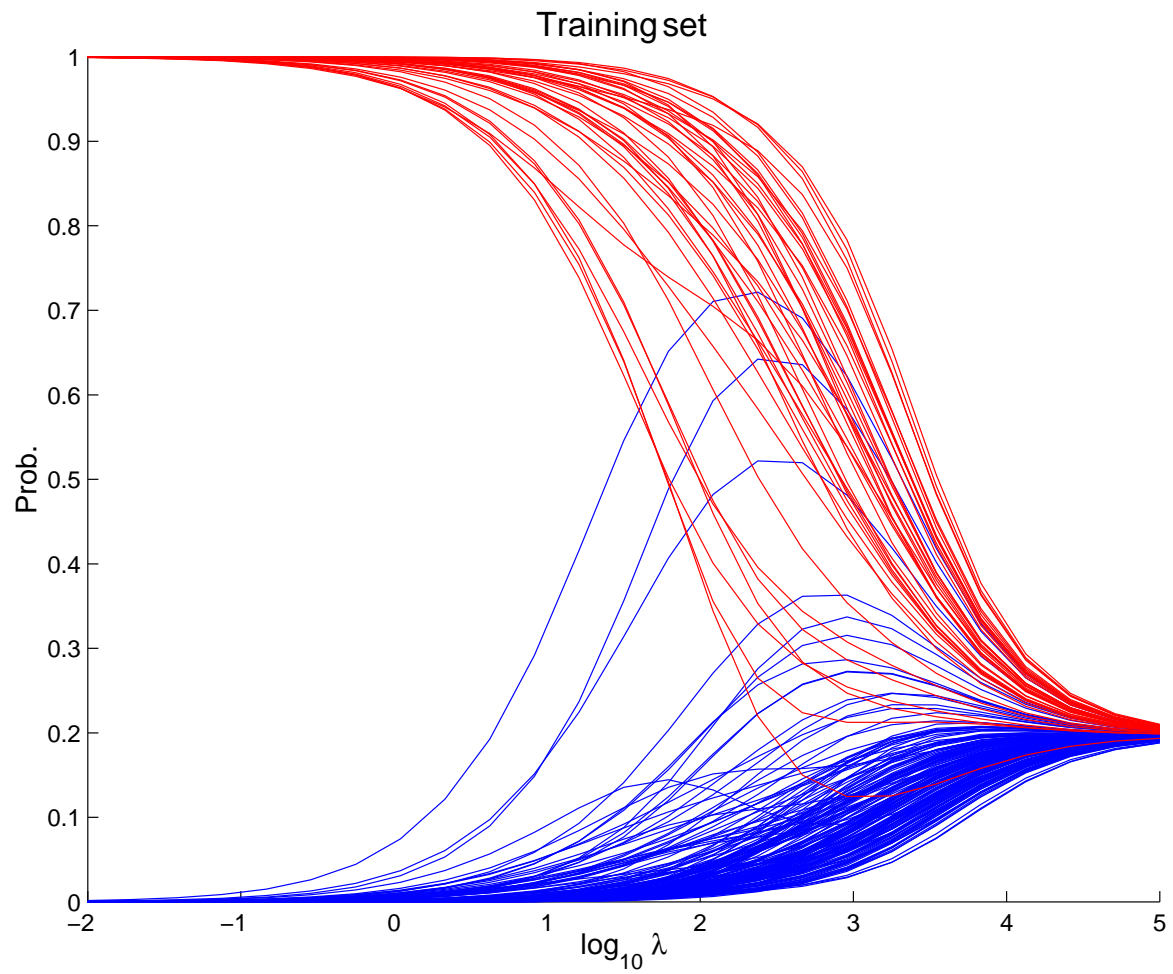


Results for Hyp50

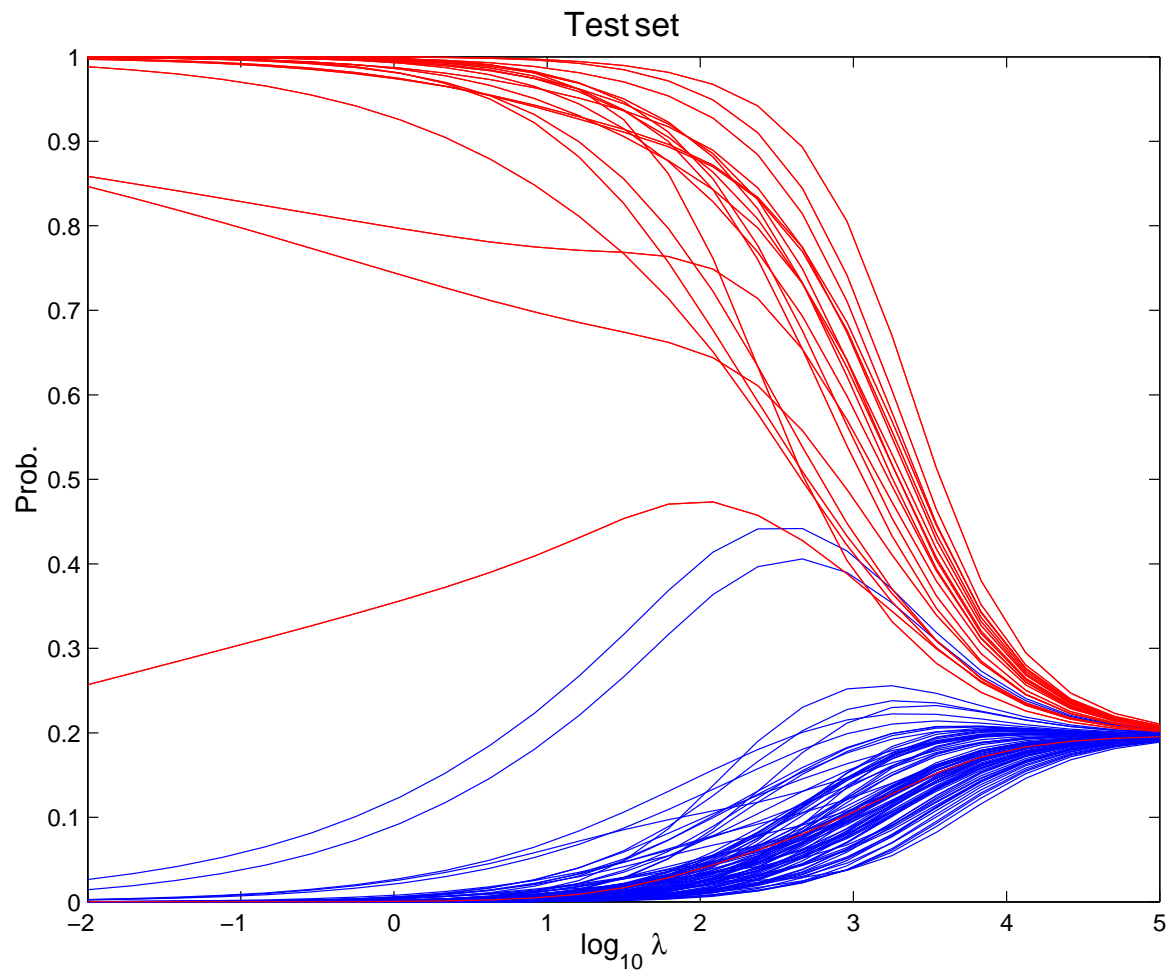
- Curve takes minimum in $\lambda = 10^{0.92} = 8.25$.
- How good is prediction?



Classification probabilities



Classification probabilities



Outlook

- Model selection: how to choose the right beta's?
- false discovery rate?

