Statistical Learning, Fall 2024-5

Class notes 6

Classification methods: linear and traditional

Recall we are in a setting where $Y \in \mathcal{G} = \{g_1, \ldots, g_K\}$ unordered set. Of special interest is the binary classification case $K = 2$. When thinking of predictive modeling in a probabilistic setting, we have repeatedly discussed it as modeling the conditional distribution $Pr(Y|X)$ or its properties like the conditional expectation $\mathbb{E}(Y|X)$.

The probabilistic approaches we have discussed so far (like least squares regression or nearest neighbors) fall under the category of *discriminative methods* that aim to directly model this conditional distribution.

In the classification context, we can consider a different approach, termed *generative modeling*, which uses Bayes rule:

$$Pr(Y|X) = \frac{Pr(X,Y)}{Pr(X)} = \frac{Pr(Y)Pr(X|Y)}{Pr(X)} \propto Pr(Y)Pr(X|Y),$$

that is, if we can estimate the two distributions on the right, we can get a good estimate of which probabilities on the left are bigger or smaller, and use that for prediction. This is especially relevant to classification, because Y takes only K values, so there is a finite set of distributions $Pr(X|Y = g_k)$ to estimate.

In the toolbox of "traditional" methods for classification, we can find methods of both types, in particular we will discuss:

- "Discriminative" methods (including non-probabilistic):
  - Linear regression for classification
  - Logistic regression (typically used for $K = 2$, often called Multinomial regression for bigger K)
  - Support vector (linear) classification
- Generative methods:
  - Linear discriminant analysis (LDA)
  - Generalizations: Quardratic (QDA) and regularized (RDA) discriminant analysis
  - Naive Bayes

Linear regression for classification

For $K = 2$ we already discussed the option of coding $Y \in \{0, 1\}$, performing a least squares regression, and if we interpret the results as:

$$\hat{f}(X) = \hat{\mathbb{X}}(Y|X) = \hat{Pr}(Y = 1|X),$$

then it's natural to use $\hat{Y} = \mathbb{I}\{\hat{f} > 0.5\}$ as the classification rule.
An important caveat is that there is no guarantee that for predictions generated this way, we get $\hat{f}(X) \in [0, 1]$, which hampers the interpretation as probabilities.

For $K > 2$, the standard approach to applying linear regression is through *one-hot encoding*, dividing the problem into K binary problems of separating each class k from all the others, and coding each one of these as a linear regression problem as above, that is denote on the training set:

$$\tilde{Y}_{ik} = \begin{cases} 1 & \text{if } Y_i = g_k \\ 0 & \text{otherwise} \end{cases},$$

and $\mathbb{Y}_{n\times K}$ is now a matrix whose i, k entry is $\tilde{Y}_{ik}$. Now we can still solve a linear regression over every column, writing it compactly as:

$$\hat{B} = \arg\min_{B\in\mathbb{R}^{p\times K}} \|\mathbb{Y} - XB\|_F^2 \text{ (Frobenius norm = sum of squares).}$$

Each column of B is simply a least squares solution for one class.

Notes:

- If the matrix $\mathbb{X}$ contains an intercept, then it is easy to see that $\sum_{k=1}^{K} \hat{f}_k(X) = 1$, and therefore it's enough to solve $K - 1$ problems, as in the case $K = 2$.
- The natural class prediction in this approach is $\hat{Y}(X) = \arg\max_k \hat{f}_k(X)$, the class with the highest score.
- A major problem with this approach is *masking*, where if one class is in the middle between two others, a prediction model may never predict the class in the middle (example on the board).

Logistic regression

We recall the general idea of generalized linear models (GLMs), that replace the standard linearity assumptions with:

$$g(\mathbb{E}(Y|X)) = X^T\beta,$$

with g prperly selected invertible function so that $g^{-1} : \mathbb{R} \to L$, where L is the set of legal values for $\mathbb{E}(Y|X)$, for example for 2-class classification where $Y|X \sim Ber(g^{-1}(X^T\beta))$, we will want $L = (0, 1)$.

GLMs have extensive theory related to exponential families, and practical applications in methods like Poisson regression (dealing with count data) and logistic regression. For our purpose we will concentrate on 2-class classification and logistic regression:

$$logit(X) = \log \frac{Pr(Y=1|X)}{1-Pr(Y=1|X)} = X^T\beta \;\; \Rightarrow \;\; Pr(Y=1|X) = \mathbb{E}(Y|X) = logit^{-1}(X^T\beta) = \frac{\exp(X^T\beta)}{1+\exp(X^T\beta)}.$$

Unlike in our discussion of linear models and least squares regression, GLMs are closely tied to probabilistic assumptions — for example, for logistic regression it is that the logit is really linear in the covariates. Given the assumptions, GLMs are fitted to data by maximizing the (log)-likelihood of the parameters β. We note that for properly defined GLMs, including Poisson and logistic regression, the likelihood is convex in β so standard second-order methods like Newton-Raphson are guaranteed to converge to the optimum (in traditional statistics, variants of this are called Fisher scoring and are described as iteratively reweighted least squares). The maximum log-likelihood problem for logistic regression (with $Y_i \in \{0,1\}$) can reformulated as:

$$\hat{\beta} = \arg\max_{\beta} \sum_{i=1}^{n} \mathbb{I}(Y_i = 1)(X_i^T\beta) - \log(1+\exp(X_i^T\beta)).$$

As noted, logistic regression requires strong assumptions about the true model. If we are willing to make these assumptions, we get some major additional benefits:

1. Statistical inference mechanism, similar to this we have for least squares linear regression (for GLMs it relies on asymptotic approximations rather than exact finite sample theory), which offers:
   - Inference on importance of predictors (Wald tests instead of t-tests in OLS)
   - Generalized ANOVA and F-tests for model selection
   - Resulting approaches for variable and model selection like forward-stepwise
2. Non-probabilistic interpretation of $X^T\hat{\beta}$ as a "scorecard" which weights the different properties of the observation in making a prediction decision.

We also note that the decision boundary $\hat{Y}_i = \mathbb{I}\left\{\hat{Pr}(Y=1|X) > 0.5\right\}$ is linear in logistic regression:

$$\hat{Pr}(Y=1|X) > 0.5 \;\Leftrightarrow\; logit^{-1}(X^T\hat{\beta}) > 0.5 \;\Leftrightarrow\; X^T\hat{\beta} > 0.$$

Logistic regression for $K > 2$ (also called multinomial regression)

Recall we now have $Y \in \mathcal{G} = \{g_1, \ldots, g_K\}$. The standard generalization assumes:

$$\begin{aligned} \log \frac{Pr(Y=g_1|X)}{Pr(Y=g_K|X)} &= X^T\beta_1 \\ &\vdots \\ \log \frac{Pr(Y=g_{K-1}|X)}{Pr(Y=g_K|X)} &= X^T\beta_{K-1} \end{aligned}$$

We can easily transform this to probabilities for the classes by multiplying and summing up:

$$\begin{aligned} Pr(Y = g_k|X) &= \exp\left(X^T\beta_k\right) Pr(Y = g_K|X),\ k < K \\ &\Rightarrow \quad Pr(Y = g_k|X) = \frac{\exp\left(X^T\beta_k\right)}{1 + \sum_{l<K} \exp\left(X^T\beta_l\right)},\ Pr(Y = g_K|X) = \frac{1}{1 + \sum_{l<K} \exp\left(X^T\beta_l\right)}. \end{aligned}$$

where the transformation uses the identity $Pr(Y = g_K|X) = \exp\left(X^T 0\right) Pr(Y = g_K|X)$ and the fact that the probabilities sum to 1.

Notes:

1. Fitting this model to data is done by maximizing the likelihood as before (we already have the probabilities written explicitly).
2. This can be interpreted as associating a vector β_k with each class, and prefering class k to l if:
$$Pr(Y = k|X) > Pr(Y = l|X) \ \Leftrightarrow\ X^T\beta_k > X^T\beta_l.$$
3. The choice of K as the "reference class" in the denominator and $\beta_K = 0$ seems arbitrary, but the solution is invariant to this choice in the sense that the predicted probabilities of the maximum likelihood solution $\hat{Pr}(Y = k|X)$ are unaffected by this choice (the coefficient vectors $\hat{\beta}_k$ *are* affected of course). Proof of this: HW3