Statistical Learning, Fall 2024-5

Class notes 7

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 properly 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:

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

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:

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

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

Linear discriminant analysis (LDA) and extensions

Recall generative methods rely on decomposing:

$$Pr(Y = g_k|X) = \frac{Pr(X|Y = g_k)Pr(Y = g_k)}{Pr(X)} := \frac{f_k(X)\pi_k}{\sum_{l=1}^{K} f_l(X)\pi_l},$$

where we use f_k to denote the conditional distribution of X for class k and π_k for the marginal probability of class k. Thus, a generative model requires:

- Modeling assumptions and a model form for f_k
- Knowledge or estimate of the marginals π_k

The key assumption of LDA is that the predictor vector $X \in \mathbb{R}^p$ is normally distributed for each class, with a common covariance matrix:

$$X|Y = g_k \sim N(\mu_k, \Sigma).$$

Hence, fitting an LDA model to data requires estimating the common covariance matrix Σ and the class centers $\mu_1, \ldots, \mu_K$.

Before discussing estimation, we can start by analyzing LDA decision boundaries between classes and proving that they are linear:

$$\begin{aligned}
f_k(X) &= \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(X-\mu_k)^T\Sigma^{-1}(X-\mu_k)\right) \\
Pr(Y = g_k|X) &> Pr(Y = g_l|X) \;\Leftrightarrow\; f_k(X)\pi_k > f_l(X)\pi_l \\
&\Leftrightarrow -0.5(X-\mu_k)^T\Sigma^{-1}(X-\mu_k) + \log(\pi_k) > -0.5(X-\mu_l)^T\Sigma^{-1}(X-\mu_l) + \log(\pi_l) \\
&\Leftrightarrow \left[-0.5\mu_k^T\Sigma^{-1}\mu_k + 0.5\mu_l^T\Sigma^{-1}\mu_l + \log(\pi_k) - \log(\pi_l)\right] + X^T\Sigma^{-1}(\mu_k - \mu_l) > 0.
\end{aligned}$$

A different perspective shows LDA classifies according to the minimal *Mahalanobis distance* from class centers (assume $\pi_k = \pi_l$ for simplicity):

$$Pr(Y = g_k|X) > Pr(Y = g_l|X) \;\Leftrightarrow\; (X-\mu_k)^T\Sigma^{-1}(X-\mu_k) < (X-\mu_l)^T\Sigma^{-1}(X-\mu_l).$$

(Drawing on the board).

If we want to fit LDA to data, we need to estimate the parameters: $\mu_1, \ldots, \mu_K$, Σ, $\pi_1, \ldots, \pi_K$. This is naturally done by "maximum likelihood" from the data:

$$\begin{aligned}
\hat{\pi}_k &= \frac{\sum_{i=1}^{n} \mathbb{I}\{Y_i = g_k\}}{n}\,,\; \hat{\mu}_k = \frac{\sum_{\mathbb{I}\{Y_i = g_k\}} X_i}{\sum_{i=1}^{n} \mathbb{I}\{Y_i = g_k\}} \\
\hat{\Sigma} &= \frac{1}{n-K}\sum_{k=1}^{K}\sum_{\mathbb{I}\{Y_i = g_k\}} (X_i - \hat{\mu}_k)(X_i - \hat{\mu}_k)^T \;\text{ (Unbiased estimate).}
\end{aligned}$$

An interesting property: if $K = 2, \pi_1 = \pi_2 = 0.5$, then the optimal decision boundary between classes is the same for LDA and least squares regression (with (0,1) coding).

Quadratic and regularized discriminant analysis (QDA, RDA)

For QDA we make a sneakily important change in the assumed model, which is now:

$$X|Y = g_k \sim N(\mu_k, \Sigma_k),$$

allowing a different covariance matrix for each class. The effect is that the decision boundaries between classes are no longer linear:

$$\begin{aligned} Pr(Y = g_k|X) &> Pr(Y = g_l|X) \Leftrightarrow f_k(X)\pi_k > f_l(X)\pi_l \\ &\Leftrightarrow -0.5(X-\mu_k)^T\Sigma_k^{-1}(X-\mu_k) + \log(\pi_k) - 0.5\log(|\Sigma_k|) \\ &\quad > -0.5(X-\mu_l)^T\Sigma_l^{-1}(X-\mu_l) + \log(\pi_l) - 0.5\log(|\Sigma_l|) \\ &\Leftrightarrow -0.5X^T(\Sigma_k^{-1} - \Sigma_l^{-1})X + X^T\left(\Sigma_k^{-1}\mu_k - \Sigma_l^{-1}\mu_l\right) + a(l,k) > 0, \end{aligned}$$

so we now get a quadratic decision function to separate classes l, k.

When estimating from data, we now need to estimate the K covariance matrices:

$$\hat{\Sigma}_k = \frac{1}{n_k - 1} \sum_{\mathbb{I}\{Y_i = g_k\}} (X_i - \hat{\mu}_k)(X_i - \hat{\mu}_k)^T \text{ (Unbiased estimate).}$$

Note, that we can define two ways to get quadratic decision boundaries:

- Run LDA with an expanded basis of size p^2: $x_1, \ldots, x_p,\ x_1^2, \ldots, x_p^2,\ x_1x_2, \ldots, x_{p-1}x_p$.
- Run QDA in the original p-dimensional space.

Are the two methods the same? Not at all, but they both yield quadratic decision boundaries. The first option does not really make sense with the probabilistic assumptions, since if x_j is Gaussian, x_j^2 or x_jx_k are not Gaussian by definition!

A problem with QDA is that it substantially increases the number of parameters: from $K + K \times p + O(p^2)$ of LDA to $K + K \times p + O(K \times p^2)$ of QDA, and of course adds flexibility. This justifies trying to regularize and obtain methods that trade-off bias and variance. One such proposal is RDA (Friedman, 1989):

$$\hat{\Sigma}_k(\alpha) = \alpha\hat{\Sigma}_k + (1-\alpha)\hat{\Sigma},$$

which combines the global estimate of covariance with the class-specific ones. This offers interesting tradeoffs between bias and variance or between linearity and non-linearity.

Naive Bayes assumption

This method was developed in completely different settings, but probabilistically we can describe it in the following generative way:

$$f_k(X) = f_{k1}(X_1) \times f_{k2}(X_2) \times \ldots \times f_{kp}(X_p),$$

that is, we assume the coordinates of X are *conditionally independent* given the class k. In the context of discriminant analysis methods, we can offer a Naive-Bayes like analogy by assuming that Σ_k is diagonal:

$$X|Y = g_k \sim N\left(\mu_k, diag(\sigma_{k1}^2, \ldots, \sigma_{kp}^2)\right),$$

this is a simplified variant of QDA, with fewer parameters and stronger (independence) assumptions.

LDA vs logistic regression

Both methods make strong probabilistic assumptions and lead to linear decision boundaries, but they are not the same. Generative methods by nature make stronger assumptions and are more powerful when the assumptions are justified (need fewer data to reach the same quality results). In the book it is illustrated that when LDA assumptions hold, LDA requires 30% less training data than logistic regression to reach the same quality of prediction!

Support vector classification (SVC)

The final "linear" method we discuss is based on a different, non-probabilistic approach. Recall, a hyper-plane L in $\mathbb{R}^p$ is defined by a linear constraint:

$$L = \left\{x \in \mathbb{R}^p : x^T\beta + \beta_0 = 0\right\},$$

with corresponding properties:

1. If $x_1, x_2 \in L$ then:

$$\beta \perp (x_1 - x_2) \quad : \quad \beta^t(x_1 - x_2) = x_1^T\beta + \beta_0 - (x_2^T\beta + \beta_0) = 0.$$

2. Signed distance of a point x from the hyper-plane L: If $\beta_0 = 0$ and $\|\beta\|_2 = 1$ then it's easy to see that the (signed) Euclidean distance is $d(x, L) = x^T\beta$, generalization for general β, β_0:

$$d(x, L) = \frac{1}{\|\beta\|_2}(x^T\beta + \beta_0).$$

Now, assume we are given a training set $T = (\mathbb{X}, \mathbb{Y})$ for a 2-class classification problem, and assume for simplicity we encode $Y_i \in \{\pm 1\}$. We define a hyperplane L as *separating* for this data if:

$$Y_i = 1 \Leftrightarrow d(X_i, L) > 0 \ , \ Y_i = -1 \Leftrightarrow d(X_i, L) < 0.$$

When the dimension p is low, such a hyperplane may not exist at all. When p is high, it is likely to exist, and then there will be many (a continuum) of such solutions.

The idea in SVC is to select the hyperplane which "best" separates the data, as defined by the minimal distance of any point to the hyperplane:

$$\max_{\beta, \beta_0} \min_i Y_i \cdot \frac{1}{\|\beta\|_2}(X^T\beta + \beta_0).$$

This problem can be formulated as equivalent optimization problems:

$$\max M \text{ s.t. } \|\beta\|_2 = 1 \ , \ Y_i \cdot (X^T\beta + \beta_0) \geq M \ \forall i$$
$$\min \|\beta\|_2^2 \text{ s.t. } Y_i \cdot (X^T\beta + \beta_0) \geq 1 \ \forall i,$$

that last formulation is the standard SVC problem as described in the literature. It leads to a linear model, where new observations are classified based on the sign of $d(x_0, L(\beta, \beta_0))$.

The term "support vector" relates to the nature of the solution to this problem:

- The optimal coefficient vector $\hat{\beta}$ has the form $\hat{\beta} = \sum_{i=1}^n \hat{\alpha}_i Y_i X_i$.
- $\hat{\alpha}_i = 0$ for all except the *support vectors* which attain equality in the constraint:

$$\hat{\alpha}_i > 0 \Rightarrow (X_i^T\hat{\beta} + \hat{\beta}_0) = 1 \ , \ \hat{\alpha}_i = 0 \Leftarrow (X_i^T\hat{\beta} + \hat{\beta}_0) > 1.$$

Since the number of support vectors is typically small, very efficient algorithms for SVC with big data (large n, large p) have been designed based on searching these support vectors, and this is some of the claim to fame of support vector methods.

What we have described so far is the "hard-margin" SVC, an extension is the soft-margin SVC, which allows violations of the constraint:

$$\min \|\beta\|_2^2 \qquad \text{s.t. } Y_i \cdot (X^T\beta + \beta_0) \geq 1 - \xi_i \ \forall i,$$
$$\xi_i \geq 0 \ \forall i \ , \ \sum_{i=1}^n \xi_i \leq C.$$

The interesting property of this problem is that with some simple manipulations and given the (Lagrange) equivalence between constrained and penalized optimization, an equivalent formulation gives a ridge-penalized problem with a nice (Hinge) loss function:

$$(\hat{\beta}, \hat{\beta}_0) = \arg\min \sum_{i=1}^n \left(1 - Y_i(X_i^T\beta + \beta_0)\right)_+ + \lambda\|\beta\|_2^2.$$

We can see the close similarity to logistic regression with ridge penalty:

$$(\hat{\beta}, \hat{\beta}_0) = \arg\min \sum_{i=1}^n \frac{1}{1 + \exp\left[-Y_i(X_i^T\beta + \beta_0)\right]} + \lambda\|\beta\|_2^2.$$

In fact, if we draw the loss functions for both problems we can see they are quite similar (in particular become parallel as the "margin" $Y_i(X_i^T\beta + \beta_0)$ becomes big).

Classification methods: summary

Ways of generating linear decision boundaries:

1. Logistic regression: maximize log likelihood with logit link, generalization to k-class
2. SVC: similar discriminative method derived from a geometric/computational perspective

3. Generative method: LDA (naturally applies to $K > 2$ classes)

Extensions and generalizations:

1. Regularized logistic regression or soft-margin SVC: add regularization to control estimation error
2. QDA and RDA: add flexibility and quadratic decision boundaries
3. Naive Bayes: different simplifying assumptions (independence given class), non-linear decision boundary but heavily regularized (few parameters)