Machine learning lecture slides

COMS 4771 Fall 2020

Classification I: Linear classification

Outline

- Logistic regression and linear classifiers
- Example: text classification
- Maximum likelihood estimation and empirical risk minimization
- Linear separators
- Surrogate loss functions

- Suppose x is given by d real-valued features, so $x \in \mathbb{R}^d$, while $y \in \{-1, +1\}$.
- *Logistic regression model* for (X, Y):
  - $Y \mid X = x$ is Bernoulli (but taking values in $\{-1, +1\}$ rather than $\{0, 1\}$) with parameter $\sigma(x^{\mathsf{T}} w) := \frac{1}{1+\exp(-x^{\mathsf{T}} w)}$.
  - *Sigmoid function* $\sigma(t) := 1/(1 + e^{-t})$
  - $w \in \mathbb{R}^d$ is parameter vector of interest
  - w not involved in marginal distribution of X (which we don't care much about)

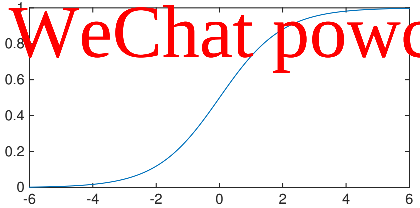

Figure 1: Logistic (sigmoid) function

Log-odds in logistic regression model

- *Sigmoid function* $\sigma(t) := 1/(1+e^{-t})$
  - Useful property: $1 - \sigma(t) = \sigma(-t)$
  - $\Pr(Y = +1 \mid X = x) = \sigma(x^\intercal w)$
  - $\Pr(Y = -1 \mid X = x) = 1 - \sigma(x^\intercal w) = \sigma(-x^\intercal w)$
  - Convenient formula: for each $y \in \{-1, +1\}$,

$$\Pr(Y = y \mid X = x) = \sigma(y x^\intercal w).$$

- Log-odds in the model is given by a linear function:

$$\ln \frac{\Pr(Y = +1 \mid X = x)}{\Pr(Y = -1 \mid X = x)} = x^\intercal w.$$

- Just like in linear regression, common to use feature expansion!
  - E.g., affine feature expansion $\varphi(x) = (1, x) \in \mathbb{R}^{d+1}$

Optimal classifier in logistic regression model

- Recall that *Bayes classifier* is

$$f^\star(x) = \begin{cases} +1 & \text{if } \Pr(Y = +1 \mid X = x) > 1/2 \\ -1 & \text{otherwise.} \end{cases}$$

- If distribution of (X, Y) comes from logistic regression model with parameter w, then Bayes classifier is

$$\begin{aligned} f^\star(x) &= \begin{cases} +1 & \text{if } x^\mathsf{T} w > 0 \\ -1 & \text{otherwise.} \end{cases} \\ &= \operatorname{sign}(x^\mathsf{T} w). \end{aligned}$$

- This is a *linear classifier*
  - Compute linear combination of features, then check if above threshold (zero)
  - With affine feature expansion, threshold can be non-zero
- Many other statistical models for classification data lead to a linear (or affine) classifier, e.g., Naive Bayes

Geometry of linear classifiers

- *Hyperplane* specified by *normal vector* $w \in \mathbb{R}^d$:
  - $H = \{x \in \mathbb{R}^d : x^\mathsf{T} w = 0\}$
  - This is the *decision boundary* of a linear classifier
  - Angle θ between x and w has

$$\cos(\theta) = \frac{x^\mathsf{T} w}{\|x\|_2 \|w\|_2}$$

  - Distance to hyperplane given by $\|x\|_2 \cdot \cos(\theta)$
  - x is on same side of H as w iff $x^\mathsf{T} w > 0$

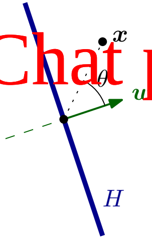

Figure 3: Decision boundary of linear classifier

- With feature expansion, can obtain other types of decision boundaries

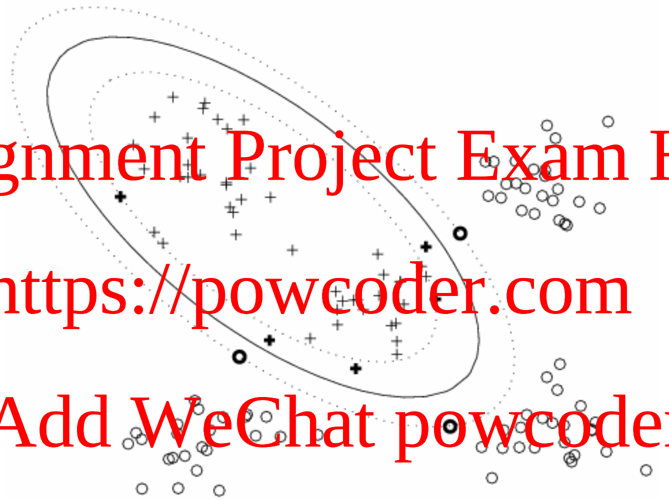

Figure 4: Decision boundary of linear classifier with quadratic feature expansion

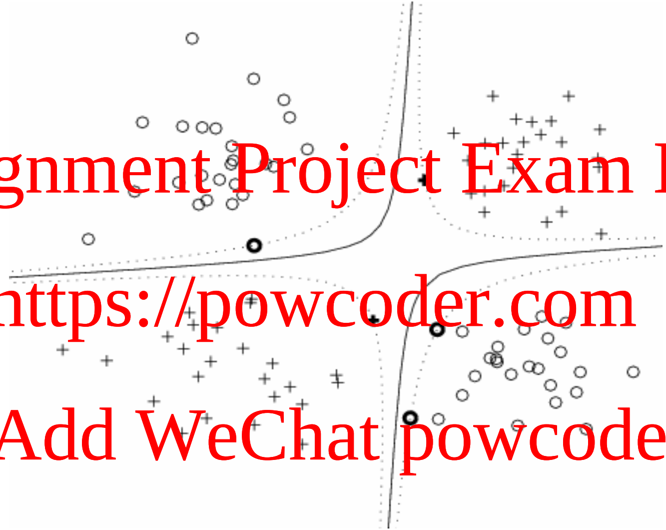

Figure 5: Decision boundary of linear classifier with quadratic feature expansion (another one)

MLE for logistic regression

- Treat training examples as iid, same distribution as test example
- Log-likelihood of w given data $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^d \times \{-1, +1\}$:

$$-\sum_{i=1}^{n} \ln(1 + \exp(-y_i x_i^{\mathsf{T}} w)) + \{\text{terms not involving } w\}$$

- No "closed form" expression for maximizer
- (Later, we'll discuss algorithms for finding approximate maximizers using iterative methods like gradient descent.)

Example: Text classification (1)

- Data: articles posted to various internet message boards
- Label: -1 for articles on "religion", $+1$ for articles on "politics"
- Features:
  - Vocabulary of $d = 61188$ words
  - Each document is a binary vector $x = (x_1, \ldots, x_d) \in \{0,1\}^d$, where $x_j = \mathbf{1}_{\{\text{document contains } j\text{-th vocabulary word}\}}$
- Logistic regression model:

$$\ln \frac{\Pr_w(Y = \text{politics} \mid X = x)}{\Pr_w(Y = \text{religion} \mid X = x)} = w^\mathsf{T} x$$

  - Each weight in weight vector $w = (w_1, \ldots, w_{61188})$ corresponds to a vocabulary word

Example: Text classification (2)

- Found $\hat{w}$ that approximately maximizes likelihood given 3028 training examples
- Test error rate on 2017 examples is about 8.5%
- Vocabulary words with 10 highest (most positive) coefficients:
  - israel, gun, government, american, news, clinton, rights, guns, israeli, politics
- Vocabulary words with 10 lowest (most negative) coefficients:
  - god, christian, bible, jesus, keith, christians, religion, church, christ, athos

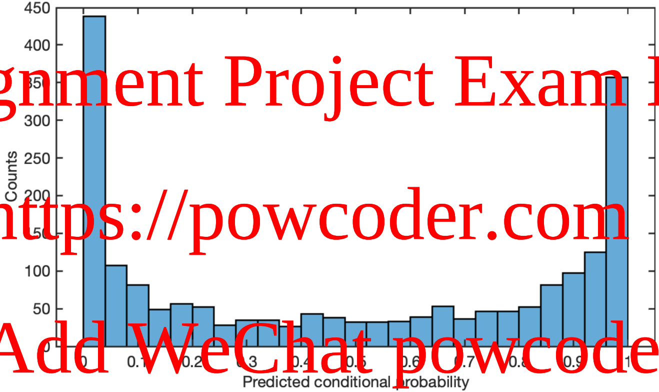

Figure 6: Histogram of $\Pr_{\hat{w}}(Y = \text{politics} \mid X = x)$ values on test data

Example: Text classification (4)

- Article with $\Pr_{\hat{w}}(Y = \text{politics} \mid X = x) \approx 0.0$:

  Rick, I think we can safely say, 1) Robert is not the only person who understands the Bible, and 2), the leadership of the LDS church historicly never has. Let's consider some "personal interpretations" and see how much trust we should put in "Orthodox Mormonism", which could never be confused with Orthodox Christianity. [...]

Example: Text classification (5)

- Article with $\Pr_{\hat{w}}(Y = \text{politics} \mid X = x) \approx 0.5$:

  Does anyone know where I can access an online copy of the proposed "jobs" or "stimulus" legislation? Please E-mail me directly and if anyone else is interested, I can post this information.

- Article with $\Pr_{\hat{w}}(Y = \text{politics} \mid X = x) \approx 1.0$:

  [A reprint of a Village Voice article by Robert I. Friedman titled "The Enemy Within" about the Anti-Defamation League.]

Zero-one loss and ERM for linear classifiers

- Recall: error rate of classifier f can also be written as risk:

$$\mathcal{R}(f) = \mathbb{E}[\mathbf{1}_{\{f(X) \neq Y\}}] = \Pr(f(X) \neq Y),$$

  where loss function is zero-one loss.
- For classification, we are ultimately interested in classifiers with small error rate

  - I.e., small (zero-one loss) risk
- Just like for linear regression, can apply plug-in principle to derive *ERM*, but now for linear classifiers.

  - Find $w \in \mathbb{R}^d$ to minimize

$$\widehat{\mathcal{R}}(w) := \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{\operatorname{sign}(x_i^{\mathsf{T}} w) \neq y_i\}}.$$

Performance of ERM for linear classifiers

- **Theorem**: In IID model, ERM solution $\hat{w}$ satisfies

$$\mathbb{E}[\mathcal{R}(\hat{w})] \leq \min_{w \in \mathbb{R}^d} \mathcal{R}(w) + O\left(\sqrt{\frac{d}{n}}\right)$$

- Inductive bias assumption: there is a linear classifier with low error rate, so $\min_{w \in \mathbb{R}^d} \mathcal{R}(w)$ is small.

- Unfortunately, solving this optimization problem, even for linear classifiers, is computationally intractable.
  - (Sharp contrast to ERM optimization problem for linear regression!)

Linearly separable data

- Training data is *linearly separable* if there exists a linear classifier with training error rate zero.
- (Special case where ERM optimization problem is tractable.)
- There exists $w \in \mathbb{R}^d$ such that $\operatorname{sign}(x_i^\top w) = y_i$ for all $i = 1, \dots, n$.
- Equivalent:

$$y_i x_i^\top w > 0 \quad \text{for all } i = 1, \dots, n$$

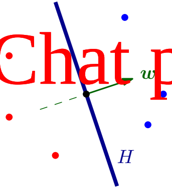

Figure 7: Linearly separable data

Figure 8: Data that is not linearly separable

Finding a linear separator I

- Suppose training data $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^d \times \{-1, +1\}$ is linearly separable.
- How to find a linear separator (assuming one exists)?
- Method 1: solve linear feasibility problem

- Method 2: approximately solve logistic regression MLE

Surrogate loss functions I

- Often, a linear separator will not exist.
- Regard each term in negative log-likelihood as *logistic loss*

$$\ell_{\text{logistic}}(s) := \ln(1 + \exp(-s))$$

- C.f. Zero-one loss: $\ell_{0/1}(s) := \mathbf{1}_{\{s \leq 0\}}$
- Scaling of ℓ_{logistic} is upper-bound on $\ell_{0/1}$: a *surrogate loss*:

$$\ell_{0/1}(s) \leq \frac{1}{\ln 2}\ell_{\text{logistic}}(s) = \ell_{\text{logistic}_2}(s).$$

- Small (empirical) ℓ_{logistic}-risk implies small (empirical) $\ell_{0/1}$-risk

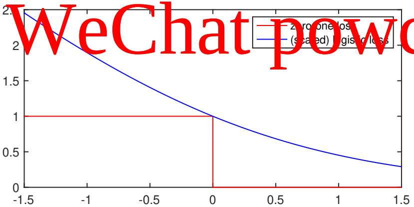

Figure 9: Comparing zero-one loss and (scaled) logistic loss

Surrogate loss functions II

- Another example: *squared loss*
  - $\ell_{\text{sq}}(s) = (1 - s)^2$
    - Note: $(1 - y_i x_i^\top w)^2 = (y_i - x_i^\top w)^2$ since $y_i \in \{-1, +1\}$
  - Weird: $\ell_{\text{sq}}(s) \to \infty$ as $s \to -\infty$.
    - Minimizing $\widehat{\mathcal{R}}_{\ell_{\text{sq}}}$ does not necessarily give a linear separator, even if one exists.
  - A fix: $\ell_{\text{msq}}(s) := \max\{0, 1 - s\}^2$ (*modified squared loss*)

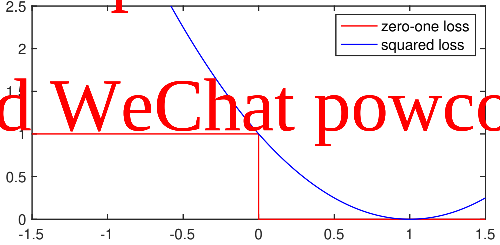

Figure 10: Comparing zero-one loss and squared loss

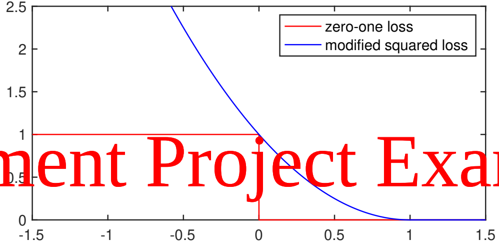

Figure 11: Comparing zero-one loss and modified squared loss

(Regularized) empirical risk minimization for classification with surrogate losses

- We can combine these surrogate losses with regularizers, just as when we discussed linear regression
- This leads to regularized ERM objectives:

$$\frac{1}{n}\sum_{i=1}^{n}\ell(y_i x_i^{\mathsf{T}} w) + \Phi(w)$$

where

- ℓ is a (surrogate) loss function
- Φ is a regularizer (e.g., $\Phi(w) = \lambda\|w\|_2^2$, $\Phi(w) = \lambda\|w\|_1$)