Machine learning lecture slides

Classification I: Linear classification

Outline

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

- 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 $\sigma(x^\top w) := \frac{1}{1+\exp(-x^\top w)}$.
  - [illegible] on t

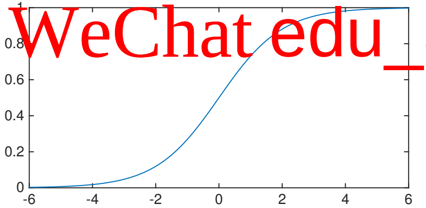


Figure 1: Logistic (sigmoid) function

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

-

$$\ln \frac{\Pr(Y = +1 \mid X}{\Pr(Y = -1 \mid X}$$

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

- Recall that *Bayes classifier* is

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

-

$$f \quad x$$

$$-1$$

$$= \operatorname{sign}(x^\top$$

- 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

- *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

$$x^{\mathsf{T}} w$$

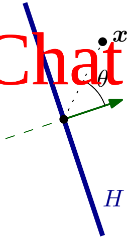


Figure 3: Decision boundary of linear classifier

- 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)

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

$w\}$

- No "closed form" expression for maximiz
- (Later, we'll discuss algorithms for findin maximizers using iterative methods like g

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

  $0, 1\}^d$,

-

$$\ln \frac{\Pr_w(Y = \text{politics} \mid}{\Pr_w(Y = \text{religion} \mid}$$

  - Each weight in weight vector w onds 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:

-

  christ, athos

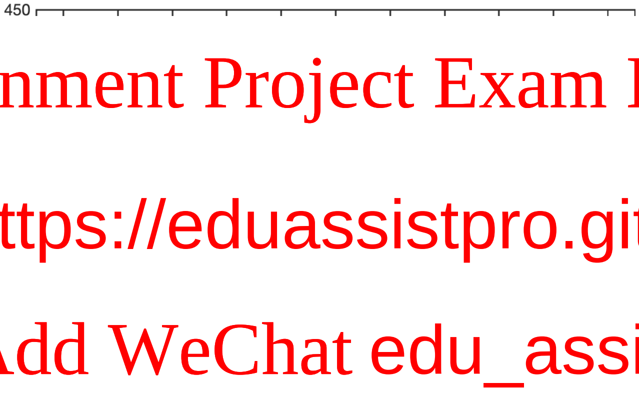


Figure 6: Histogram of $\Pr_{\hat{w}}(Y = \text{politics} \mid X = x)$ values on test data

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

- 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

-

  titled "The Enemy Within" about the Anti-
  League.]

- 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.

-

- Just like for linear regression, can apply plug
derive *ERM*, but now for linear classifiers
  - Find $w \in \mathbb{R}^d$ to minimize

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

Performance of ERM for linear classifiers

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

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

-

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

- 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 $\operatorname{sign}(x_i^{\intercal} w) = y_i$ for all $i = 1, \ldots, n$
-

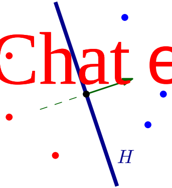


Figure 7: Linearly separable data

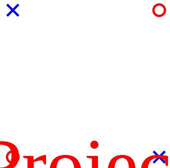


Figure 8: Data that is not linearly separable

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

- 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: $\ell_{0/1}(s) := \mathbf{1}_{s \le 0}$
- *oss*:
- ℓ_{logistic} $\ell_{0/1}$-risk

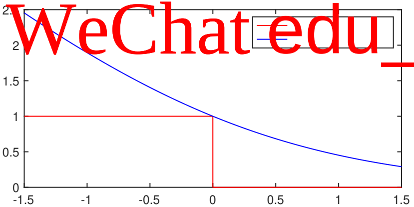


Figure 9: Comparing zero-one loss and (scaled) logistic loss

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 $\mathcal{R}_{\ell_{\text{sq}}}$ does not necessarily give a linear separator,

)

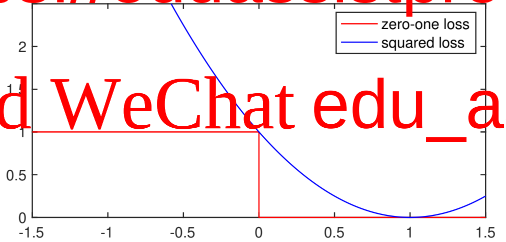
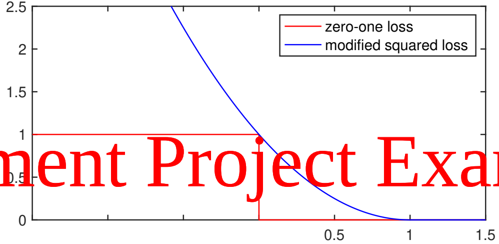


Figure 10: Comparing zero-one loss and squared loss


zero-one loss
modified squared loss
2.5
2
1.5
1
0
0.5
1
1.5

(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:

where

- ℓ is a (surrogate) loss function
- Φ is a regularizer (e.g., $\Phi(w) =$