Slides adapted from Prof Carpuat and Duraiswami

CMSC 422 Introduction to Machine Learning

Lecture 15 A Probabilistic View of Machine Learning I

Furong Huang / furongh@cs.umd.edu

UNIVERSITY OF MARYLAND

Approximating the 0-1 loss with surrogate loss functions

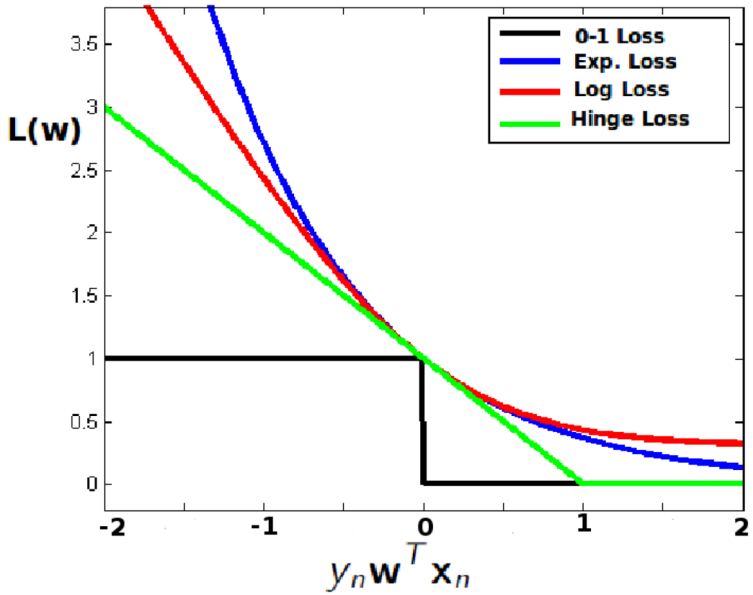
Examples (with b = 0)

Hinge loss $[1 - y_n \mathbf{w}^T \mathbf{x}_n]_+ = \max\{0, 1 - y_n \mathbf{w}^T \mathbf{x}_n\}$

Log loss $\log[1 + \exp(-y_n \mathbf{w}^T \mathbf{x}_n)]$

Exponential loss $\exp(-y_n \mathbf{w}^T \mathbf{x}_n)$

What if $b \neq 0$?



Example: subgradient of hinge loss

For a given example n

$$\partial_w \max\{0, 1 - y_n(\boldsymbol{w} \cdot \boldsymbol{x}_n + b)\} \tag{6.22}$$

$$= \partial_w \begin{cases} 0 & \text{if } y_n(\boldsymbol{w} \cdot \boldsymbol{x}_n + b) > 1 \\ y_n(\boldsymbol{w} \cdot \boldsymbol{x}_n + b) & \text{otherwise} \end{cases} \tag{6.23}$$

$$= \begin{cases} \partial_w 0 & \text{if } y_n(\boldsymbol{w} \cdot \boldsymbol{x}_n + b) > 1 \\ \partial_w y_n(\boldsymbol{w} \cdot \boldsymbol{x}_n + b) & \text{otherwise} \end{cases} \tag{6.24}$$

$$= \begin{cases} \mathbf{0} & \text{if } y_n(\boldsymbol{w} \cdot \boldsymbol{x}_n + b) > 1 \\ y_n \boldsymbol{x}_n & \text{otherwise} \end{cases} \tag{6.25}$$

Subgradient Descent for Hinge Loss

Algorithm 23 HINGEREGULARIZEDGD($\mathbf{D}$, λ, $MaxIter$)

1: $\boldsymbol{w} \leftarrow \langle 0, 0, \ldots 0 \rangle$, $b \leftarrow 0$ // initialize weights and bias
2: **for** $iter = 1 \ldots MaxIter$ **do**
3: $\boldsymbol{g} \leftarrow \langle 0, 0, \ldots 0 \rangle$, $g \leftarrow 0$ // initialize gradient of weights and bias
4: **for all** $(\boldsymbol{x}, y) \in \mathbf{D}$ **do**
5: **if** $y(\boldsymbol{w} \cdot \boldsymbol{x} + b) \leq 1$ **then**
6: $\boldsymbol{g} \leftarrow \boldsymbol{g} + y\,\boldsymbol{x}$ // update weight gradient
7: $g \leftarrow g + y$ // update bias derivative
8: **end if**
9: **end for**
10: $\boldsymbol{g} \leftarrow \boldsymbol{g} - \lambda \boldsymbol{w}$ // add in regularization term
11: $\boldsymbol{w} \leftarrow \boldsymbol{w} + \eta \boldsymbol{g}$ // update weights
12: $b \leftarrow b + \eta g$ // update bias
13: **end for**
14: **return** $\boldsymbol{w}, b$

What is the perceptron optimizing?

Algorithm 5 PERCEPTRONTRAIN(**D**, *MaxIter*)

```
w_d ← 0, for all d = 1 ... D                          // initialize weights
b ← 0                                                  // initialize bias
for iter = 1 ... MaxIter do
    for all (x,y) ∈ D do
        a ← Σ_{d=1}^{D} w_d x_d + b                    // compute activation for this example
        if ya ≤ 0 then
            w_d ← w_d + y x_d, for all d = 1 ... D     // update weights
            b ← b + y                                  // update bias
        end if
    end for
end for
return w_0, w_1, ..., w_D, b
```

Loss function is a variant of the hinge loss

$$\max\{0, -y_n(\mathbf{w}^T\mathbf{x}_n + b)\}$$

Recap: Linear Models

Lets us separate model definition from training algorithm **(Gradient Descent)**

Summary

Gradient descent

- A generic algorithm to minimize objective functions
- Works well as long as functions are well behaved (ie convex)
- Subgradient descent can be used at points where derivative is not defined
- Choice of step size is important

Optional: can we do better?

- For some objectives, we can find closed form solutions (see CIML 6.6)

Today's topics

- Bayes rule review
- A probabilistic view of machine learning
  - Joint Distributions
  - Bayes optimal classifier
- Statistical Estimation
  - Maximum likelihood estimates
  - Derive relative frequency as the solution to a constrained optimization problem

Bayes Rule

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$ Bayes' rule

we call P(A) the "prior"

and P(A|B) the "posterior"

Bayes, Thomas (1763) An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London,* **53:370-418**

…by no means merely a curious speculation in the doctrine of chances, but necessary to be solved in order to a sure foundation for all our reasonings concerning past facts, and what is likely to be hereafter…. necessary to be considered by any that would give a clear account of the strength of *analogical* or *inductive reasoning…*

Exercise: Applying Bayes Rule

Consider the 2 random variables

- A = You have the flu
- B = You just coughed

Assume

- P(A) = 0.05
- P(B|A) = 0.8
- P(B|not A) = 0.2

What is P(A|B)?

Answer

- Via Logic
  - ✓ Assume 100 students – 5 have the flu. 80% (4) of the students who have the flu cough; 20% (19) of the students who don't have the flu cough; So the chance that you have the flu is 4/23
- Via Bayes Rule
  - ✓ $P(A|B)P(B)=P(B|A)P(A)$.
  - ✓ $P(B)=0.8*0.05+0.2*(1-0.05)=0.04+0.19=0.23$
  - ✓ $P(A|B)=0.8*0.05/0.23 =0.04/0.23=4/23$

Using a Joint Distribution

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

Using a Joint Distribution

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

Given the joint distribution, we can find the probability of any logical expression E involving these variables

$$P(E) = \sum_{\text{rows matching } E} P(\text{row})$$

Using a Joint Distribution

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

Given the joint distribution, we can make inferences

- E.g., P(Male | Poor)?
- Or P(Wealth | Gender, Hours)?

Recall: Machine Learning as Function Approximation

Problem setting

- Set of possible instances X
- Unknown target function $f: X \rightarrow Y$
- Set of function hypotheses $H = \{h \mid h: X \rightarrow Y\}$

Input

- Training examples $\{(x^{(1)}, y^{(1)}), \ldots (x^{(N)}, y^{(N)})\}$ of unknown target function f

Output

- Hypothesis $h \in H$ that best approximates target function f

Recall: Formal Definition of Binary Classification (from CIML)

TASK: BINARY CLASSIFICATION

Given:

1. An input space $\mathcal{X}$
2. An unknown distribution $\mathcal{D}$ over $\mathcal{X} \times \{-1, +1\}$

Compute: A function f minimizing: $\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[f(\boldsymbol{x}) \neq y\right]$

The Bayes Optimal Classifier

Assume we know the data generating distribution $\mathcal{D}$

We define the **Bayes Optimal classifier** as

$$f^{(\mathrm{BO})}(\hat{x}) = \arg\max_{\hat{y} \in \mathcal{Y}} \mathcal{D}(\hat{x}, \hat{y})$$

Theorem: Of all possible classifiers, the Bayes Optimal classifier achieves the smallest zero/one loss

Bayes error rate

- Defined as the error rate of the Bayes optimal classifier
- Best error rate we can ever hope to achieve under zero/one loss

The Bayes Optimal Classifier

Assume we know the data generating distribution $\mathcal{D}$

We define the **Bayes Optimal classifier** as

$$f^{(\mathrm{BO})}(\hat{x}) = \arg\max_{\hat{y}} \mathcal{D}(\hat{x}, \hat{y})$$

The fier
achi

Bay
D
B

If we had access to $\mathcal{D}$,
Finding an optimal classifier would be trivial!

we don't have access to $\mathcal{D}$
So let's try to estimate it instead!

What does "training" mean in probabilistic settings?

- Training = estimating $\mathcal{D}$ from a finite training set
  - We typically assume that $\mathcal{D}$ comes from a specific family of probability distributions
  - e.g., Bernouilli, Gaussian, etc
  - Learning means inferring parameters of that distributions
  - e.g., mean and covariance of the Gaussian

Training assumption: training examples are iid

- **Independently and Identically distributed**
  - i.e. as we draw a sequence of examples from $\mathcal{D}$, the n-th draw is independent from the previous n-1 sample

- This assumption is usually false!
  - But sufficiently close to true to be useful

How can we estimate the joint probability distribution from data?

What are the challenges?

Maximum Likelihood Estimation

- Find the parameters that maximize the probability of the data

- Example: how to model a biased coin?

  (on board)

Maximum Likelihood Estimation

- Example: how to model a k-sided die? (on board)

Today's topics

Bayes rule review

A probabilistic view of machine learning

- Joint Distributions
- Bayes optimal classifier

Statistical Estimation

- Maximum likelihood estimates
- Derive relative frequency as the solution to a constrained optimization problem

UNIVERSITY OF
MARYLAND

Furong Huang
3251 A.V. Williams, College Park, MD 20740
301.405.8010 / furongh@cs.umd.edu