

Assignment Project Exam Help

Machine learning lecture slides

COMS 4771 Fall 2020

<https://powcoder.com>

Add WeChat powcoder

Classification III: Classification objectives

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

- ▶ Scoring functions
- ▶ Cost-sensitive classification
- ▶ Conditional probability estimation
- ▶ Reducing multi-class to binary
- ▶ Fairness in classification

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

Scoring functions in general

- ▶ Statistical model: $(X, Y) \sim P$ for distribution P over $\mathcal{X} \times \{-1, +1\}$

- ▶ Binary classifiers are generally of the form

$$x \mapsto \text{sign}(h(x))$$

for some scoring function $h: \mathcal{X} \rightarrow \mathbb{R}$.

- ▶ E.g. Bayes classifier uses scoring function $h(x) = \eta(x) - 1/2$ where $\eta(x) = \Pr(Y = +1 \mid X = x)$
- ▶ Use with loss functions like $\ell_{0/1}$, ℓ_{logistic} , ℓ_{sq} , ℓ_{msq} , ℓ_{hinge}

$$\mathcal{R}(h) = \mathbb{E}[\ell(Yh(X))]$$

- ▶ Issues to consider:
 - ▶ Different types of mistakes have different costs
 - ▶ How to get $\Pr(Y = +1 \mid X = x)$ from $h(x)$?
 - ▶ More than two classes

Cost-sensitive classification

- Cost matrix for different kinds of mistakes (for $c \in [0, 1]$)

	$\hat{y} = -1$	$\hat{y} = +1$
$y = -1$	0	c
$y = +1$	$1 - c$	0

(Why can we restrict attention to $c \in [0, 1]$?)

- Cost-sensitive ℓ -loss:

$$\ell^{(c)}(y, \hat{y}) = (\mathbf{1}_{\{y=+1\}} \cdot (1 - c) + \mathbf{1}_{\{y=-1\}} \cdot c) \cdot \ell(y\hat{y}).$$

► If ℓ is convex in \hat{y} , then so is $\ell^{(c)}(y, \cdot)$

- Cost-sensitive (empirical) risk:

$$\mathcal{R}^{(c)}(h) := \mathbb{E}[\ell^{(c)}(Y, h(X))]$$

$$\widehat{\mathcal{R}}^{(c)}(h) := \frac{1}{n} \sum_{i=1}^n \ell^{(c)}(y_i, h(x_i))$$

Minimizing cost-sensitive risk

- ▶ What is the analogue of Bayes classifier for cost-sensitive (zero-one loss) risk?

- ▶ Let $\eta(x) = \Pr(Y=+1 | X=x)$
- ▶ Fix x ; what is conditional cost-sensitive risk of predicting \hat{y} ?

$$\eta(x) \cdot (1 - c) \cdot \mathbf{1}_{\{\hat{y}=-1\}} + (1 - \eta(x)) \cdot c \cdot \mathbf{1}_{\{\hat{y}=+1\}}.$$

- ▶ Minimized when

$$\hat{y} = \begin{cases} +1 & \text{if } \eta(x) \cdot (1 - c) > (1 - \eta(x)) \cdot c \\ -1 & \text{otherwise} \end{cases}$$

- ▶ So use scoring function $h(x) = \eta(x) - c$
 - ▶ Equivalently, use η as scoring function, but threshold at c instead of $1/2$
- ▶ Where does c come from?

Example: balanced error rate

- Balanced error rate: $\text{BER} := \frac{1}{2}\text{FNR} + \frac{1}{2}\text{FPR}$
- Which cost sensitive risk to try to minimize?

$\frac{1}{2} \cdot \text{BER}$

$$= \Pr(h(X) \leq 0 \mid Y = +1) + \Pr(h(X) > 0 \mid Y = -1)$$

$$= \frac{1}{2} \cdot \Pr(h(X) \leq 0 \wedge Y = +1) + \frac{1}{2} \cdot \Pr(h(X) > 0 \wedge Y = -1)$$

where $\pi = \Pr(Y = +1)$.

- Therefore, we want to use the following cost matrix:

	$\hat{y} = -1$	$\hat{y} = +1$
$y = -1$	0	$\frac{1}{1-\pi}$
$y = +1$	$\frac{1}{\pi}$	0

- This corresponds to $c = \pi$.

Importance-weighted risk

- ▶ Perhaps the world tells you how important each example is
- ▶ Statistical model: $(X, Y, W) \sim P$
 - ▶ W is (non-negative) importance weight of example (X, Y)
- ▶ Importance-weighted ℓ -risk of h :

$$\mathbb{E}[W \cdot \ell(Yh(X))]$$

- ▶ Estimate from data $(x_1, y_1, w_1), \dots, (x_n, y_n, w_n)$:

$$\frac{1}{n} \sum_{i=1}^n w_i \cdot \ell(y_i h(x_i))$$

Add WeChat powcoder

Conditional probability estimation (1)

- ▶ How to get estimate of $\eta(x) = \Pr(Y = +1 \mid X = x)$?
- ▶ Useful if want to know expected cost of a prediction

Assignment Project Exam Help

$$\mathbb{E}[\ell_{0/1}^{(c)}(Yh(X)) \mid X = x] = \begin{cases} (1 - c) \cdot \eta(x) & \text{if } h(x) \leq 0 \\ c \cdot (1 - \eta(x)) & \text{if } h(x) > 0 \end{cases}$$

- ▶ Squared loss risk minimized by scoring function

$$h(x) = 2\eta(x) - 1.$$

Therefore, given h , can estimate η using $\hat{\eta}(x) = \frac{h(x)}{2}$

- ▶ Recipe:
 - ▶ Find scoring function h that (approximately) minimizes (empirical) squared loss risk
 - ▶ Construct conditional probability estimate $\hat{\eta}$ using above formula

Conditional probability estimation (2)

- ▶ Similar strategy available for logistic loss
- ▶ But not for hinge loss!

Assignment Project Exam Help

- ▶ Hinge loss risk is minimized by $\hat{h}(x) = \text{sgn}(2\hat{\eta}(x) - 1)$
 - ▶ Cannot recover η from h
- ▶ Caveat: If using insufficiently expressive functions for h (e.g., linear functions), may be far from minimizing squared loss risk
 - ▶ Fix: use more flexible models (e.g., feature expansion)

<https://powcoder.com>

Add WeChat powcoder

Application: Reducing multi-class to binary

- ▶ Multi-class: Conditional probability function is vector-valued function

$$\eta(x) = \begin{bmatrix} \Pr(Y = 1 | X = x) \\ \vdots \\ \Pr(Y = K | X = x) \end{bmatrix}$$

- ▶ Reduction: learn K scalar-valued functions, the k -th function is supposed to approximate

$$\eta_k(x) = \Pr(Y = k | X = x).$$

- ▶ This can be done by create K binary classification problems, where in problem k , label is $\mathbf{1}_{\{y=k\}}$.
- ▶ Given the K learned conditional probability functions $\hat{\eta}_1, \dots, \hat{\eta}_K$, we form a final predictor \hat{f}

$$\hat{f}(x) = \arg \max_{k=1, \dots, K} \hat{\eta}_k(x).$$

When does one-against-all work well?

- ▶ If learned conditional probability functions $\hat{\eta}_k$ are accurate, then behavior of one-against-all classifier \hat{f} is similar to optimal classifier

$$f^*(x) = \arg \max_{k=1, \dots, K} \Pr(Y = k \mid X = x).$$

- ▶ **Claim:** <https://powcoder.com>

$$\text{err}(\hat{f}) \leq \text{err}(f^*) + 2 \cdot \mathbb{E}[\max_k |\hat{\eta}_k(X) - \eta_k(X)|].$$

Add WeChat powcoder

- ▶ Use of predictive models (e.g., in admissions, hiring, criminal justice) has raised concerns about whether they offer “fair treatment” to individuals and/or groups
 - ▶ We will focus on group-based fairness
 - ▶ Individual-based fairness also important, but not as well-studied

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

Disparate treatment

- ▶ Often predictive models work better for some groups than for others

▶ Example: face recognition (Buolamwini and Gebru, 2018; Lohr, 2018)

Assignment Project Exam Help

Color Matters in Computer Vision

Facial recognition algorithms made by Microsoft, IBM and Face++ were more likely to misidentify the gender of black women than white men.



Gender was misidentified in up to 1 percent of lighter-skinned males in a set of 385 photos.

Gender was misidentified in up to 1 percent of darker-skinned males in a set of 271 photos.



Gender was misidentified in up to 7 percent of lighter-skinned females in a set of 296 photos.

Gender was misidentified in 35 percent of darker-skinned females in a set of 271 photos.

Possible causes of unfairness

- ▶ People deliberately being unfair
- ▶ Disparity in number of available training data for different groups
- ▶ Disparity in usefulness of available features for different groups
- ▶ Disparity in relevance of prediction problem for different groups
- ▶ ...

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

- ▶ ProPublica (investigative journalism group) studied a particular predictive model being used to determine “pre-trial detention”

- ▶ Angwin et al., 2016

- ▶ Judge needs to decide whether or not an arrested defendant should be released while awaiting trial

- ▶ Predictive model (“COMPAS”) provides an estimate of

$\Pr(Y = 1 \mid X = x)$ where

$Y = 1_{\{\text{will commit (violent) crime if released}\}}$ and X is “features” of defendant.

- ▶ Study argued that COMPAS treated black defendants unfairly in a certain sense

- ▶ What sense? How do they make this argument?

- ▶ Setup:
 - ▶ X : features for individual
 - ▶ A : group membership attribute (e.g., race, sex, age, religion)
 - ▶ Y : outcome variable to predict (e.g., "will repay loan", "will re-offend")
 - ▶ \hat{Y} : prediction of outcome variable (as function of (X, A))
 - ▶ For simplicity, assume A , Y , and \hat{Y} are $\{0, 1\}$ -valued
- ▶ Many fairness criteria are based on joint distribution of

$$(A, Y, \hat{Y})$$

- ▶ Caveat: Often, we don't have access to Y in training data

Classification parity

- Fairness criterion: Classification parity

Assignment Project Exam Help

- Sounds reasonable, but easy to satisfy with perverse methods
- Example: trying to predict $Y = \mathbf{1}_{\{\text{will repay loan if given one}\}}$
- Suppose conditional distributions of (\hat{Y}, Y) given A are as follows.

$(A = 0)$			$(A = 1)$		
$\hat{Y} = 0$	$\hat{Y} = 1$		$\hat{Y} = 0$	$\hat{Y} = 1$	
$Y = 0$	1/2	0	$Y = 0$	1/4	1/4
$Y = 1$	0	1/2	$Y = 1$	1/4	1/4

- For $A = 0$ people, correctly give loans to people who will repay
- For $A = 1$ people, give loans randomly (Bernoulli(1/2))
- Satisfies criterion, but bad for $A = 1$ people

Equalized odds (1)

- Fairness criterion: Equalized odds

Assignment Project Exam Help

$\Pr(\hat{Y} = 1 | Y = y, A = 0) \approx \Pr(\hat{Y} = 1 | Y = y, A = 1)$
for both $y \in \{0, 1\}$.

- In particular, FPR and FNR must be (approximately) same across groups.

► Could also just ask for Equalized FPR, or Equalized FNR

- Previous example fails to satisfy equalized odds:

Add WeChat powcoder

$(A = 0)$			$(A = 1)$		
$\hat{Y} = 0$	$\hat{Y} = 1$		$\hat{Y} = 0$	$\hat{Y} = 1$	
$Y = 0$	1/2	0	1/4	1/4	
$Y = 1$	0	1/2	1/4	1/4	

E.g., $A = 0$ group has 0% FPR, while $A = 1$ has 50% FPR.

- Criteria imply constraints on the classifier / scoring function
 - Can try to enforce constraint during training

Equalized odds (2)

- ▶ ProPublica study:

- ▶ Found that FPR for $A = 0$ group (black defendants; 45%) was higher than FPR for $A = 1$ group (white defendants; 23%)

$(A = 0)$	$\hat{Y} = 0$	$\hat{Y} = 1$	$(A = 1)$	$\hat{Y} = 0$	$\hat{Y} = 1$
$Y = 0$	0.27	0.22	$Y = 0$	0.46	0.14
$Y = 1$	0.14	0.37	$Y = 1$	0.19	0.21

<https://powcoder.com>

Add WeChat powcoder