Prediction theory

COMS 4771 Fall 2019

Overview

- Statistical model for classification problems
- Plug-in principle
- Statistical models and MLE
- Error estimation and evaluation

Statistical model for binary outcomes

Figure 1: Coin toss

- Physical model: hard
- Statistical model: outcome is random
  - *Bernoulli distribution* with heads probability θ
  - Written as $\mathrm{Bern}(\theta)$
- Goal: correctly predict outcome

Learning to make predictions

- If θ known:

- If θ unknown:

Plug-in principle

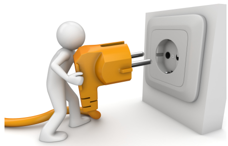


Figure 2: Plug-in

- *Plug-in principle*:
  - Estimate unknown(s) based on data (e.g., θ)
  - Plug estimates into formula for optimal prediction
- When can we estimate the unknowns?
  - Observed data should be related to the outcome we want to predict
  - *IID model*: Observations & outcome are *iid* random variables

Statistical models

- *Parametric statistical model* $\{P_\theta : \theta \in \Theta\}$
  - collection of parameterized probability distributions for observed data
- E.g., distributions on n binary outcomes treated as iid Bernoulli random variables
  - $\Theta =$
  - $P_\theta(y_1, \ldots, y_n) =$

Maximum likelihood estimation

- *Likelihood* of parameter θ (given observed data)
  - $L(\theta) = P_\theta(y_1, \ldots, y_n)$
- *Maximum likelihood estimation*: choose θ with highest likelihood
- Log-likelihood
  - E.g., $\ln L(\theta) =$
- Maximizer:

Performance of plug-in prediction I

- $\hat{\theta}$ is MLE estimate of θ from data $y_1, \ldots, y_n$
- Plug-in prediction of outcome: $\hat{y} = \mathbb{1}_{\{\hat{\theta} > 1/2\}}$
- Is this any good? Study behavior in IID model
  - $Y_1, \ldots, Y_n, Y$ are iid Bernoulli with parameter θ
  - $\hat{Y}$ is plug-in prediction

Performance of plug-in prediction II

- **Theorem**: $\Pr(\hat{Y} \neq Y) \leq \min\{\theta, 1-\theta\} + |2\theta - 1| \cdot e^{-2n(\theta-0.5)^2}$

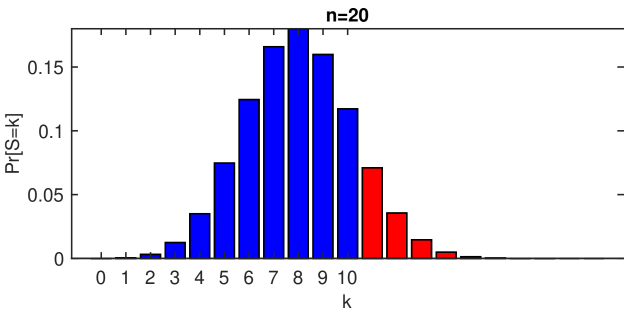


Figure 3: $n = 20$

Performance of plug-in prediction III

- **Theorem**: $\Pr(\hat{Y} \neq Y) \leq \min\{\theta, 1-\theta\} + |2\theta - 1| \cdot e^{-2n(\theta-0.5)^2}$

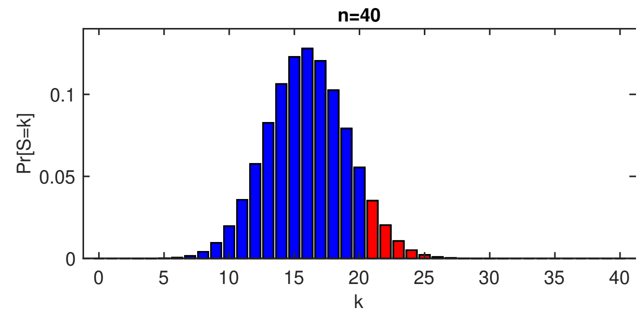


Figure 4: $n = 40$

Performance of plug-in prediction IV

- **Theorem**: $\Pr(\hat{Y} \neq Y) \leq \min\{\theta, 1-\theta\} + |2\theta - 1| \cdot e^{-2n(\theta-0.5)^2}$

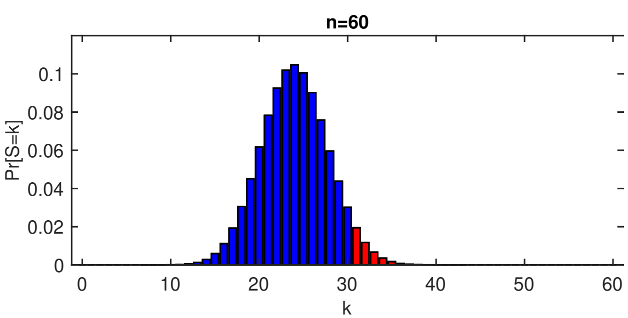


Figure 5: $n = 60$

Performance of plug-in prediction V

- **Theorem**: $\Pr(\hat{Y} \neq Y) \leq \min\{\theta, 1-\theta\} + |2\theta - 1| \cdot e^{-2n(\theta-0.5)^2}$

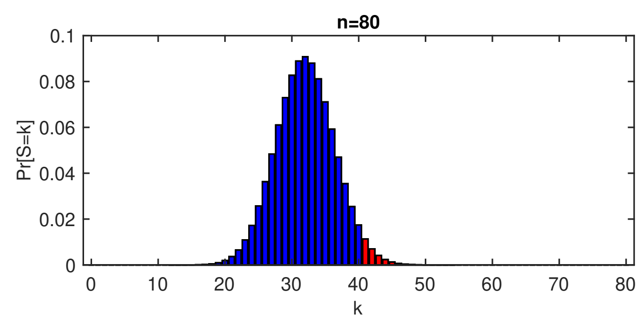


Figure 6: $n = 80$

Statistical model for labeled examples

- Example: spam filtering
- Labeled example: $(x, y) \in \mathcal{X} \times \{0, 1\}$
- $\mathcal{X}$ is input (feature) space; $\{0, 1\}$ is the output (label) space
  - $\mathcal{X}$ is not necessarily the space of inputs itself (e.g., space of all emails), but rather the space of what we measure about inputs
- We only see x, and then must make prediction of y
- Statistical model: (X, Y) is random
  - X has some *marginal probability distribution*
  - *Conditional probability distribution* of Y given $X = x$ is Bernoulli with heads probability $\eta(x)$
  - $\eta \colon \mathcal{X} \to [0, 1]$ is a function, sometimes called the *regression function*

Conditional expectations

- Consider any random variables A and B.
- Conditional expectation of A given B:
  - Written $\mathbb{E}[A \mid B]$
  - A random variable! What is its expectation?
  - *Law of iterated expectations*:

Bayes classifier

- *Optimal classifier* (*Bayes classifier*):

$$f^\star(x) = \mathbb{1}_{\{\eta(x) > 1/2\}},$$

  where η is the regression function
  - Classifier with smallest probability of mistake
  - Depends on the regression function η, which is typically unknown!
- *Optimal error rate* (*Bayes error rate*):
  - Write error rate as $\Pr(f^\star(X) \neq Y) = \mathbb{E}[\mathbb{1}_{\{f^\star(X) \neq Y\}}]$
  - In terms of η:

Example: spam filtering

- Suppose input x is a single (binary) feature, "is email all-caps?"
- How to interpret "the probability that email is spam given $x = 1$?"
- What does it mean for the Bayes classifier $f^\star$ to be optimal?

Learning prediction functions

- What to do if η is unknown?
  - Training data: $(x_1, y_1), \ldots, (x_n, y_n)$
  - Data are related to what we want to predict
  - IID model: $(X_1, Y_1), \ldots, (X_n, Y_n), (X, Y)$ are iid random variables
  - (X, Y) is the "test" example
  - (Technically, each labeled example is a $(\mathcal{X} \times \{0,1\})$-valued random variable. If $\mathcal{X} = \mathbb{R}^d$, can regard as vector of $d+1$ random variables.)

Performance of nearest neighbor classifiers

- Study in context of IID model
- Assume $\eta(\boldsymbol{x}) \approx \eta(\boldsymbol{x}')$ whenever $\boldsymbol{x}$ and $\boldsymbol{x}'$ are close.
- Let $(\boldsymbol{X}, Y)$ be the "test" example, and suppose $(\boldsymbol{X}^*, Y^*)$ is the nearest neighbor among training data.

Performance of decision trees

- Hard to analyze in the IID model!
- Simpler algorithm: assume partitioning of $\mathcal{X} = \mathbb{R}^d$ is fixed in advance before seeing any training data
- Fix leaf node, and consider training examples that reach that node.

Test error rate

- How to estimate error rate?
- IID model: Training examples $((X_i, Y_i))_{i=1}^n$ and test examples $((X'_i, Y'_i))_{i=1}^m$ are iid
- Classifier $\hat{f}$ is based only on training examples; hence, it is independent of test examples
- Conditional distribution of

$$\sum_{i=1}^{m} \mathbb{1}_{\{\hat{f}(X'_i) \neq Y'_i\}}$$

  given $((X_i, Y_i))_{i=1}^n$ and $\hat{f}$:
  - *Binomial distribution* with m trials and heads probability equal to error rate ε of $\hat{f}$
  - Written as $Z \sim \mathrm{Binom}(m, \varepsilon)$

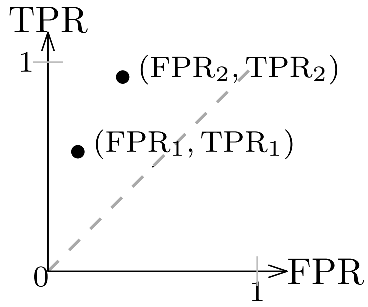
Confusion tables

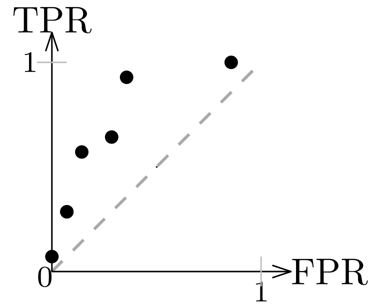
- *True positive rate* (*recall*): $\Pr(\hat{f}(X) = 1 \mid Y = 1)$
- *False positive rate*: $\Pr(\hat{f}(X) = 1 \mid Y = 0)$
- *Precision*: $\Pr(Y = 1 \mid \hat{f}(X) = 1)$
- ...
- *Confusion table*

	$\hat{y} = 0$	$\hat{y} = 1$
$y = 0$	# true negatives	# false positives
$y = 1$	# false negatives	# true positives

ROC curves

- *Receiver operating characteristic (ROC) curve*
  - What points are achievable on the TPR-FPR plane?
  - Use randomization to combine classifiers





More than two outcomes

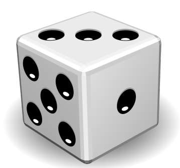


Figure 7: Six-sided die

- What if $K > 2$ possible outcomes?
- Replace coin with K-sided die
- Say Y has a *categorical distribution* over $[K] := \{1, \dots, K\}$, determined probability vector $\theta = (\theta_1, \dots, \theta_K)$
  - $\theta_k \geq 0$ for all $k \in [K]$, and $\sum_{k=1}^{K} \theta_k = 1$
  - $\Pr(Y = k) = \theta_k$
- Optimal prediction of Y if θ is known

$$\hat{y} := \arg\max_{k \in [K]} \theta_k$$

Statistical model for multi-class classification

- Statistical model for labeled examples (X, Y), where Y takes values in $[K]$
  - Now, $Y \mid X = x$ has a categorical distribution with parameter vector $\eta(x) = (\eta(x)_1, \dots, \eta(x)_K)$
  - *Conditional probability function* $\eta(x)_k := \Pr(Y = k \mid X = x)$
  - *Optimal classifier*: $f^\star(x) = \arg\max_{k \in [K]} \eta(x)_k$
  - *Optimal error rate*: $\Pr(f^\star(X) \neq Y) =$