

# Assignment Project Exam Help

Machine learning lecture slides

COMS 4771 Fall 2020

<https://powcoder.com>

Add WeChat powcoder

Prediction theory

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

- ▶ Statistical model for binary outcomes
- ▶ Plug-in principle and IID model
- ▶ Maximum likelihood estimation
- ▶ Statistical model for binary classification
- ▶ Analysis of nearest neighbor classifier
- ▶ Estimating the error rate of a classifier
- ▶ Beyond binary classification and the IID model

Assignment Project Exam Help  
<https://powcoder.com>

Add WeChat powcoder

## Statistical model for binary outcomes

- ▶ Example: coin toss
- ▶ Physical model: hard
- ▶ Statistical model: outcome is random
  - ▶ Bernoulli distribution with heads probability  $\theta \in [0, 1]$ 
    - ▶ Encode heads as 1 and tails as 0
    - ▶ Written as  $\text{Bernoulli}(\theta)$
    - ▶ Notation:  $Y \sim \text{Bernoulli}(\theta)$  means  $Y$  is a random variable with distribution  $\text{Bernoulli}(\theta)$ .
- ▶ Goal: correctly predict outcome

Add WeChat powcoder

# Optimal prediction

- ▶ Suppose  $Y \sim \text{Bernoulli}(\theta)$ .

- ▶ Suppose  $\theta$  known.

- ▶ Optimal prediction:

$$\mathbf{1}_{\{\theta > 1/2\}}$$

- ▶ Indicator function notation:

$$\mathbf{1}_{\{Q\}} = \begin{cases} 1 & \text{if } Q \text{ is true} \\ 0 & \text{if } Q \text{ is false} \end{cases}$$

- ▶ The optimal prediction is incorrect with probability

$$\min\{\theta, 1 - \theta\}$$

- ▶ If  $\theta$  unknown:

- ▶ Assume we have data: outcomes of previous coin tosses

- ▶ Data should be related to what we want to predict: same coin is being tossed

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

## Plug-in principle and IID model

- ▶ Plug-in principle:

- ▶ Estimate unknown(s) based on data (e.g.,  $\theta$ )

- ▶ 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 & (unseen) outcome are iid random variables

- ▶ iid: independent and identically distributed

- ▶ Crucial modeling assumption that makes learning possible

- ▶ When is the IID assumption not reasonable? ...

- ▶ Parametric statistical model  $\{P_\theta : \theta \in \Theta\}$ 
  - ▶ collection of parameterized probability distributions for data
  - ▶  $\Theta$  is the parameter space
  - ▶ One distribution per parameter value  $\theta \in \Theta$
- ▶ E.g., distributions on  $n$  binary outcomes treated as iid Bernoulli random variables
  - ▶  $\Theta = [0, 1]$
  - ▶ Overload notation:  $P_\theta$  is the probability mass function (pmf) for the distribution.
  - ▶ What is formula for  $P_\theta(y_1, \dots, y_n)$  for  $(y_1, \dots, y_n) \in \{0, 1\}^n$ ?

<https://powcoder.com>

Add WeChat powcoder



## Maximum likelihood estimation (1)

- ▶ Likelihood of parameter  $\theta$  (given observed data)

- ▶  $L(\theta) = P_{\theta}(y_1, \dots, y_n)$

- ▶ Maximum likelihood estimator:

- ▶ Choose  $\theta$  with highest likelihood

- ▶ Log-likelihood

- ▶ Sometimes more convenient

- ▶ It is increasing, so  $\ln L(\theta)$  orders the parameters in the same way as  $L(\theta)$

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

## Maximum likelihood estimation (2)

- ▶ Coin toss example
  - ▶ Log-likelihood

Assignment Project Exam Help

$$\ln L(\theta) = \sum_{i=1}^n y_i \ln \theta + (1 - y_i) \ln(1 - \theta)$$

- ▶ Use calculus to determine formula for maximizer
- ▶ This is a little annoying, but someone else has already done it for you:

$$\hat{\theta}_{MLE} := \frac{1}{n} \sum_{i=1}^n y_i.$$

Add WeChat powcoder

## Back to plug-in principle

- ▶ We are given data  $y_1, \dots, y_n \in \{0, 1\}^n$ , which we model using the IID model from before
- ▶ Obtain estimate  $\hat{\theta}_{\text{MLE}}$  of known  $\theta$  based on  $y_1, \dots, y_n$
- ▶ Plug-in  $\hat{\theta}_{\text{MLE}}$  for  $\theta$  in formula for optimal prediction:

$$\hat{Y} := \mathbf{1}_{\{\hat{\theta}_{\text{MLE}} > 1/2\}}.$$

<https://powcoder.com>

Add WeChat powcoder

## Analysis of the plug-in prediction (1)

- ▶ How good is the plug-in prediction?
  - ▶ Study behavior under the IID model, where  $Y_1, \dots, Y_n, Y \sim_{\text{iid}} \text{Bernoulli}(\theta)$ .
    - ▶  $Y_1, \dots, Y_n$  are the data we collected
    - ▶  $Y$  is the outcome to predict
    - ▶  $\theta$  is the unknown parameter
  - ▶ Recall: optimal prediction is incorrect with probability  $\min\{\theta, 1 - \theta\}$
  - ▶ We cannot hope  $\hat{Y}$  to beat this, but we can hope it is not much worse.

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

## Analysis of the plug-in prediction (2)

► **Theorem:**

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

► The first term is the optimal error probability.

► The second term comes from the probability that the  $\hat{\theta}_{\text{MLE}}$  is on the opposite side of  $1/2$  as  $\theta$ .

► This probability is very small when  $n$  is large!

► If  $S$  is number of heads in  $n$  independent tosses of coin with bias  $\theta$ , then  $S \sim \text{Binomial}(n, \theta)$  ([Binomial distribution](#))

Add WeChat powcoder

Assignment Project Exam Help

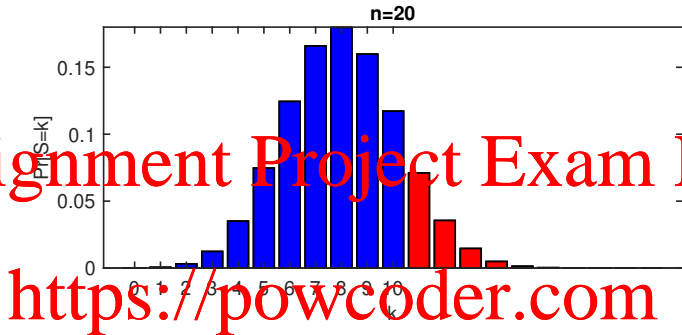


Figure 1:  $\Pr(S > n/2)$  for  $S \sim \text{Binomial}(n, \theta)$ ,  $n = 20$

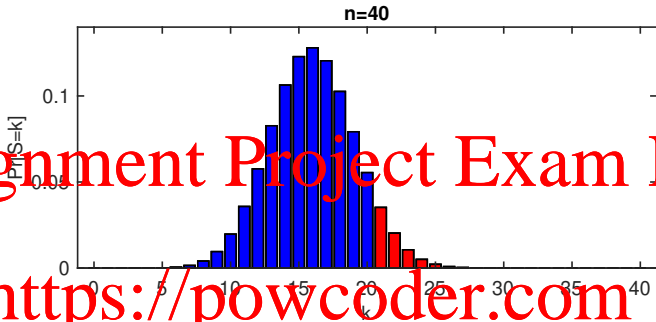
Add WeChat powcoder

Assignment Project Exam Help

<https://powcoder.com>

Figure 2:  $\Pr(S > n/2)$  for  $S \sim \text{Binomial}(n, \theta)$ ,  $n = 40$

Add WeChat powcoder

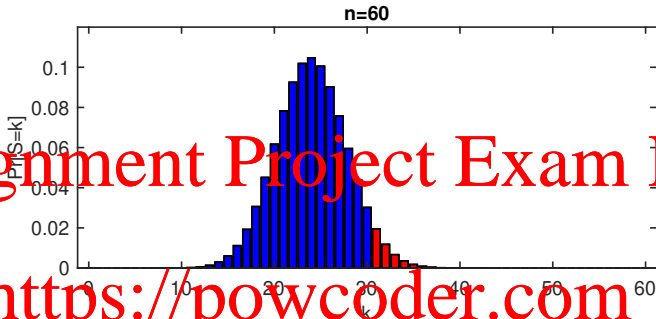


Assignment Project Exam Help

<https://powcoder.com>

Figure 3:  $\Pr(S > n/2)$  for  $S \sim \text{Binomial}(n, \theta)$ ,  $n = 60$

Add WeChat powcoder





Assignment Project Exam Help

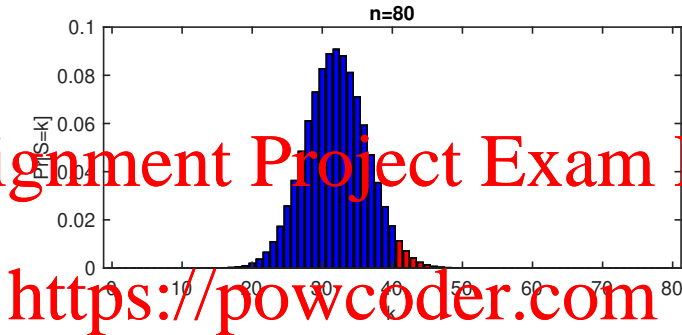


Figure 4:  $\Pr(S > n/2)$  for  $S \sim \text{Binomial}(n, \theta)$ ,  $n = 80$

Add WeChat powcoder

# Statistical model for labeled data in binary classification

- ▶ 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  $x$  (email), and then must make prediction of  $y$  (spam or not-spam)
- ▶ 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  $p(x)$
  - ▶  $\eta: \mathcal{X} \rightarrow [0, 1]$  is a function, sometimes called the regression function or conditional mean function (since  $\mathbb{E}[Y | X = x] = \eta(x)$ ).

## Error rate of a classifier

- ▶ For a classifier  $f: \mathcal{X} \rightarrow \{0, 1\}$ , the error rate of  $f$  (with respect to the distribution of  $(X, Y)$ ) is

Assignment Project Exam Help

$$\text{err}(f) := \Pr(f(X) \neq Y).$$

Recall that we had previously used the notation

<https://powcoder.com>

$$\text{err}(f, S) = \frac{1}{|S|} \sum_{(x,y) \in S} \mathbf{1}_{\{f(x) \neq y\}},$$

which is the same as  $\Pr(f(X) \neq Y)$  when the distribution of  $(X, Y)$  is uniform over the labeled examples in  $S$ .

- ▶ Caution: This notation  $\text{err}(f)$  does not make explicit the dependence on (the distribution of) the random example  $(X, Y)$ . You will need to determine this from context.

## Conditional expectations (1)

- ▶ Consider any random variables  $A$  and  $B$ .
- ▶ Conditional expectation of  $A$  given  $B$ :
  - ▶ Write  $\mathbb{E}[A | B]$
  - ▶ A random variable! What is its expectation?
  - ▶ Law of iterated expectations (a.k.a. tower property):

$\mathbb{E}[\mathbb{E}[A | B]] = \mathbb{E}[A]$   
<https://powcoder.com>

Add WeChat powcoder

## Conditional expectations (2)

- ▶ Example: roll a fair 6-sided die
  - ▶  $A$  = number shown facing up
  - ▶  $B$  = parity of number shown facing up
  - ▶  $C := \mathbb{E}[A \mid B]$  is random variable with

$$\Pr\left(C = \mathbb{E}[A \mid B = \text{odd}] = \frac{1}{3}(1 + 3 + 5) = 3\right) = \frac{1}{2}$$
$$\Pr\left(C = \mathbb{E}[A \mid B = \text{even}] = \frac{1}{3}(2 + 4 + 6) = 4\right) = \frac{1}{2}$$

Add WeChat powcoder

- Optimal classifier (Bayes classifier):

# Assignment Project Exam Help

where  $\eta$  is the conditional mean function

- Classifier with smallest probability of mistake
- Depends on the function  $\eta$ , which is typically unknown!
- Optimal error rate (Bayes error rate):
  - Write error rate as  $\text{err}(f^*) = \Pr(f^*(X) \neq Y) = \mathbb{E}[\mathbf{1}_{\{f^*(X) \neq Y\}}]$
  - Conditional on  $X$ , probability of mistake is  $\min\{\eta(X), 1 - \eta(X)\}$ .
  - So, optimal error rate is

$$\begin{aligned}\text{err}(f^*) &= \mathbb{E}[\mathbf{1}_{\{f^*(X) \neq Y\}}] \\ &= \mathbb{E}[\mathbb{E}[\mathbf{1}_{\{f^*(X) \neq Y\}} \mid X]] \\ &= \mathbb{E}[\min\{\eta(X), 1 - \eta(X)\}].\end{aligned}$$

## 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  $f^*$  to be optimal?

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

- ▶ What to do if  $\eta$  is unknown?
  - ▶ Training data:  $(x_1, y_1), \dots, (x_n, y_n)$
  - ▶ Assume data are related to what we want to predict
  - ▶ Let  $Z := (X, Y)$ , and  $Z_i := (X_i, Y_i)$  for  $i = 1, \dots, n$ .
  - ▶ IID model:  $Z_1, \dots, Z_n, Z$  are iid random variables
  - ▶  $Z = (X, Y)$  is the (unseen) “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.)

Add WeChat powcoder



# Performance of nearest neighbor classifier

- ▶ Study in context of IID model
- ▶ Assume  $\eta(x) \approx \eta(x')$  whenever  $x$  and  $x'$  are close.
  - ▶ This is where the modeling assumption comes in (via choice of distance function)!

- ▶ Let  $(X, Y)$  be the “test” example, and suppose  $(X_i, Y_i)$  is the nearest neighbor among training data

$$S = ((X_1, Y_1), \dots, (X_n, Y_n)).$$

- ▶ For large  $n$ ,  $X$  and  $X_i$  likely to be close enough so that  $\eta(X) \approx \eta(X_i)$ .

- ▶ Prediction is  $Y_i$ , true label is  $Y$ .

- ▶ Conditional on  $X$  and  $X_i$  what is probability that  $Y \neq Y_i$ ?

$$\eta(X)(1 - \eta(X_i)) + (1 - \eta(X))\eta(X_i) \approx 2\eta(X)(1 - \eta(X))$$

- ▶ Conclusion: expected error rate is

$$\mathbb{E}[\text{err}(\text{NN}_S)] \approx 2 \cdot \mathbb{E}[\eta(X)(1 - \eta(X))] \text{ for large } n$$

- ▶ Recall that optimal is  $\mathbb{E}[\min\{\eta(X), 1 - \eta(X)\}]$ .
- ▶ So  $\mathbb{E}[\text{err}(\text{NN}_S)]$  is at most twice optimal.
- ▶ Never exactly optimal unless  $\eta(x) \in \{0, 1\}$  for all  $x$ .

## Test error rate (1)

- ▶ How to estimate error rate?
- ▶ IID model:

$(X_1, Y_1), \dots, (X_n, Y_n), (X'_1, Y'_1), \dots, (X'_m, Y'_m), (X, Y)$  are iid

- ▶ Training examples (that you have):  $(X_1, Y_1), \dots, (X_n, Y_n)$
- ▶ Test examples (that you have):  $(X'_1, Y'_1), \dots, (X'_m, Y'_m)$
- ▶ Test example (that you don't have) used to define error rate:

<https://powcoder.com>

- ▶ Classifier  $\hat{f}$  is based only on training examples
- ▶ Hence, **test examples are independent of  $\hat{f}$**  (very important!)
- ▶ We would like to estimate  $\text{err}(\hat{f})$

Add WeChat powcoder

- ▶ Caution: since  $\hat{f}$  depends on training data, it is random!
- ▶ Convention: When we write  $\text{err}(\hat{f})$  where  $\hat{f}$  is random, we really mean  $\Pr(\hat{f}(X) \neq Y \mid \hat{f})$ .
- ▶ Therefore  $\text{err}(\hat{f})$  is a random variable!

## Test error rate (2)

- ▶ Conditional distribution of  $S := \sum_{i=1}^m \mathbf{1}_{\{\hat{f}(X'_i) \neq Y'_i\}}$  given training data:

▶  $S$  | training data  $\sim \text{Binomial}(m, \varepsilon)$  where  $\varepsilon := \text{err}(\hat{f})$

- ▶ By law of large numbers,

$$\frac{1}{m}S \rightarrow \varepsilon$$

as  $m \rightarrow \infty$

- ▶ Therefore, test error rate

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

is close to  $\varepsilon$  when  $m$  is large

- ▶ How accurate is the estimate? Depends on the (conditional) variance!

- ▶  $\text{var}(\frac{1}{m}S \mid \text{training data}) = \frac{\varepsilon(1-\varepsilon)}{m}$

- ▶ Standard deviation is  $\sqrt{\frac{\varepsilon(1-\varepsilon)}{m}}$

## Confusion tables

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

<https://powcoder.com>

	$f(x) = 0$	$f(x) = 1$
$y = 0$	# true negatives	# false positives
$y = 1$	# false negatives	# true positives

Add WeChat powcoder

- ▶ Receiver operating characteristic (ROC) curve

- ▶ What points are achievable on the TPR-FPR plane?

- ▶ Use randomization to combine classifiers

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

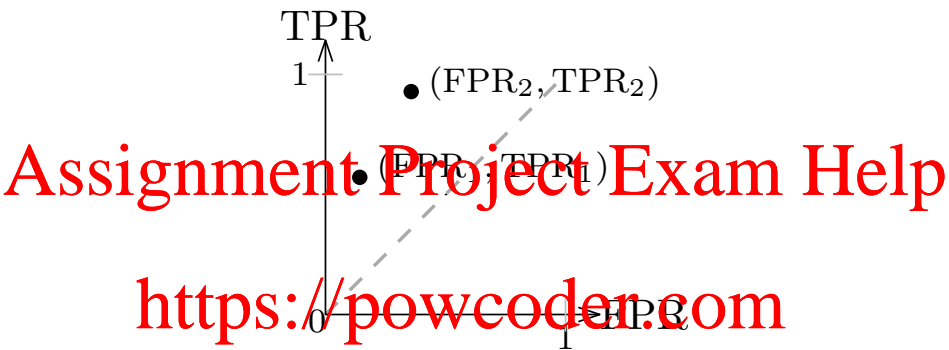


Figure 5: TPR vs FPR plot with two points

Add WeChat powcoder

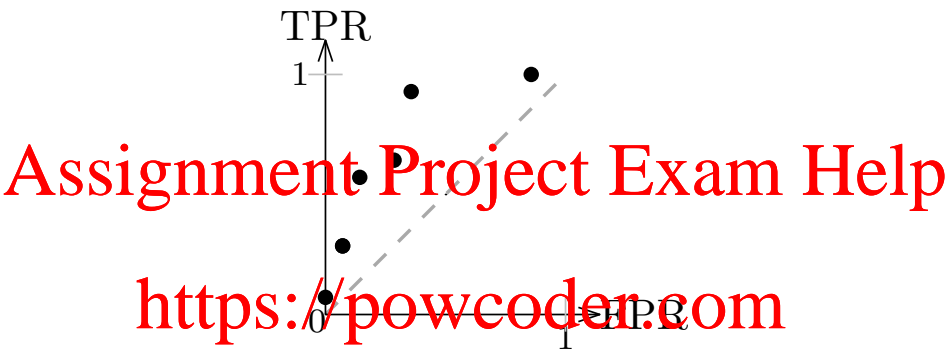


Figure 6: TPR vs FPR plot with many points

Add WeChat powcoder

## More than two outcomes

- ▶ What if there are  $K > 2$  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  $\theta$  is known

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

Add WeChat powcoder



# Statistical model for multi-class classification

- ▶ Statistical model for labeled examples  $(X, Y)$ , where  $Y$  takes values in  $[K]$

▶ Now,  $Y | X = x$  has a categorical distribution with parameter vector  $\eta(x) = (\eta(x)_1, \dots, \eta(x)_K)$

- ▶ Conditional probability function:  $\eta(x)_k := \Pr(Y = k | X = x)$

- ▶ Optimal classifier:  $f^*(x) = \arg \max_{k \in [K]} \eta(x)_k$

- ▶ Optimal error rate:  $\Pr(f^*(X) \neq Y) = 1 - \mathbb{E}[\max_k \eta(X)_k]$

Assignment Project Exam Help  
<https://powcoder.com>

Add WeChat powcoder

## Potential downsides of the IID model

- ▶ Example: Train OCR digit classifier using data from Alice's handwriting, but eventually use on digits written by Bob.

▶ What is a better evaluation?

Assignment Project Exam Help

<https://powcoder.com>

- ▶ What if we want to eventually use on digits written by both Alice and Bob?

Add WeChat powcoder