Machine learning lecture slides

Prediction theory

Outline

- Statistical model for binary outcomes
- Plug-in principle and IID model
- Maximum likelihood estimation
- Statistical model for binary classification

Statistical model for binary outcomes

- Example: coin toss
- Physical model: hard
- Statistical model: outcome is random
  - *Bernoulli distribution* with heads probability $\theta \in [0, 1]$

ith

- Goal: correctly predict outcome

Optimal prediction

- Suppose $Y \sim \mathrm{Bernoulli}(\theta)$.
  - Suppose θ known.
  - Optimal prediction:

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

  - The optimal prediction is incorrect with

$$\min\{\theta,$$

Learning to make predictions

- If θ 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

Plug-in principle and IID model

- *Plug-in principle*:
  - Estimate unknown(s) based on data (e.g., θ)
  - Plug estimates into formula for optimal prediction

-

  - *IID model*: Observations & (unseen) o m variables
  - *iid*: independent and identically distrib
  - Crucial modeling assumption that ma

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

Statistical models

- *Parametric statistical model* $\{P_\theta : \theta \in \Theta\}$
  - collection of parameterized probability distributions for data
  - Θ is the *parameter space*
  - One distribution per parameter value $\theta \in \Theta$
- li

  (*pmf*)

  for the distribution.
  - What is formula for $P_\theta(y_1, \dots, y$ $\}^n$?

Maximum likelihood estimation (1)

- *Likelihood* of parameter θ (given observed data)
  - $L(\theta) = P_\theta(y_1, \ldots, y_n)$
- *Maximum likelihood estimation*:
  - Choose θ with highest likelihood
-

Maximum likelihood estimation (2)

- Coin toss example
  - Log-likelihood

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

$$\hat{\theta}_{\mathrm{MLE}} := -$$

Back to plug-in principle

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

Analysis of the plug-in prediction (1)

- How good is the plug-in prediction?
  - Study behavior under the IID model, where $Y_1, \ldots, Y_N, Y \sim_{\text{iid}} \text{Bernoulli}(\theta)$.
    - $Y_1, \ldots, Y_n$ are the data we collected
    - Y is the outcome to predict

Y

worse.

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 $\hat{\theta}_{\mathrm{MLE}}$ is on the opposite side of $1/2$ as θ.

[illegible] n) th

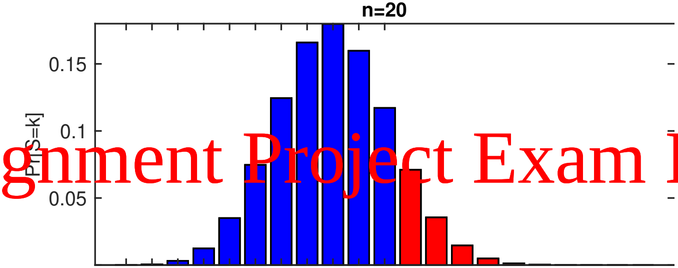

Figure 1: $\Pr(S > n/2)$ for $S \sim$

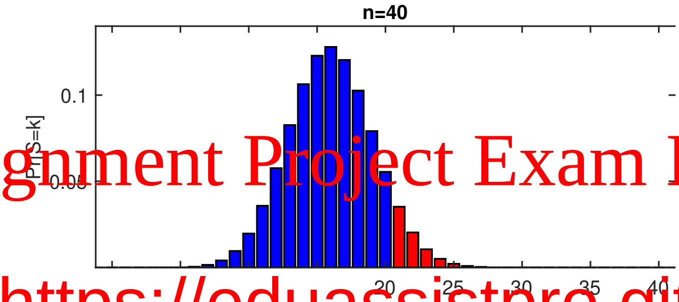

Figure 2: $\Pr(S > n/2)$ for $S \sim$

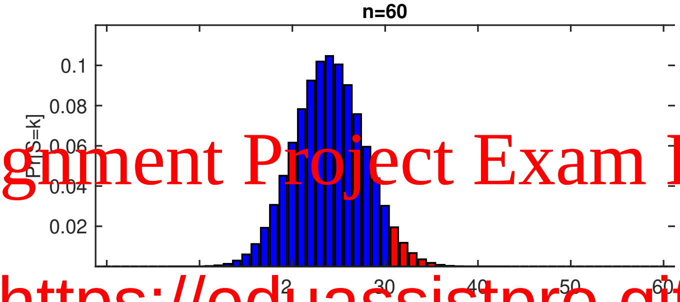

Figure 3: $\Pr(S > n/2)$ for $S \sim$

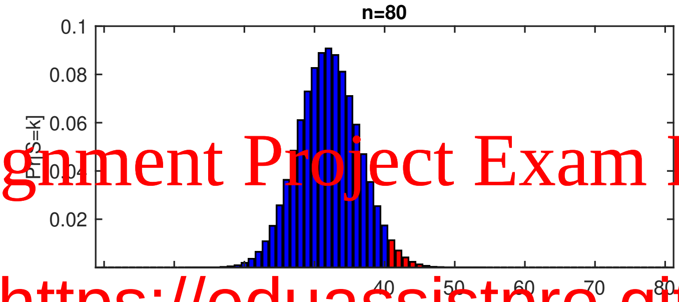

Figure 4: $\Pr(S > n/2)$ for $S \sim$

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

- y

-
  - X has some *marginal probability d*
  - *Conditional probability distribution* Bernoulli with heads probability
  - $\eta \colon \mathcal{X} \to [0, 1]$ is a function, sometim *regression function* or *conditional mean function* (since $\mathbb{E}[Y \mid X = x] = \eta(x)$).

Error rate of a classifier

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

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

$$\overline{\quad} \; |\;|_{(x,y} \quad /$$

  which is the same as $\Pr(f(X) \neq$
  (X,Y) is uniform over the labeled exampl

- Caution: This notation $\operatorname{err}(f)$ 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:
  - Written $\mathbb{E}[A \mid B]$
  - A random variable! What is its expectation?
  - *Law of iterated expectations* (a.k.a. *tower property*):

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

$$\frac{\quad}{\quad} \qquad = \frac{1}{2}$$

$$\frac{\quad}{\quad} \qquad = \frac{1}{2}$$

Bayes classifier

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

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

  where η is the conditional mean function

-
  - Write error rate as $\mathrm{err}(f^\star) = \Pr$ $\star$ $/$ $/Y\}]$
  - Conditional on X, probability of mis $\min\{\eta(X), 1 - \eta(X)\}$.
  - So, optimal error rate is

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

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 $f^\star$ to be optimal?

Learning prediction functions

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

alued

$+1$

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 $(X_{\hat{i}}, Y_{\hat{i}})$ is the

- $\eta(X) \approx \eta(X_{\hat{i}})$.
- Prediction is $Y_{\hat{i}}$, true label is Y.
- Conditional on X and $X_{\hat{i}}$, what is pro ?
  - $\eta(X)(1 - \eta(X_{\hat{i}})) + (1 - \eta(X))$

- Conclusion: expected error rate is $\mathbb{E}[\mathrm{err}(\mathrm{NN}_S)] \approx 2 \cdot \mathbb{E}[\eta(X)(1 - \eta(X))]$ for large n
  - Recall that optimal is $\mathbb{E}[\min\{\eta(X), 1 - \eta(X)\}]$.
  - So $\mathbb{E}[\mathrm{err}(\mathrm{NN}_S)]$ 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:
  $(X_1, Y_1), \ldots, (X_n, Y_n), (X'_1, Y'_1), \ldots, (X'_m, Y'_m), (X, Y)$ are iid
  - Training examples (that you have): $(X_1, Y_1), \ldots, (X_n, Y_n)$
    $'_m)$

-
- Hence, **test examples are independe**
  important!)
- We would like to estimate $\mathrm{err}(\hat{f})$
  - Caution: since $\hat{f}$ depends on traini
  - Convention: When we write $\mathrm{err}(\hat{f})$ where $\hat{f}$ is random, we really mean $\Pr(\hat{f}(X) \neq Y \mid \hat{f})$.
  - Therefore $\mathrm{err}(\hat{f})$ 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:
  - $S \mid \text{training data} \sim \text{Binomial}(m, \varepsilon)$ where $\varepsilon := \text{err}(\hat{f})$
  - By law of large numbers,

$$\frac{1}{m}\sum_{i=1}^{m}\mathbf{1}_{\{\hat{f}(X_i') \neq Y_i'\}}$$

    is close to ε 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*: $\Pr(Y = 1 \mid f(X) = 1)$
- ...
-

$y = 1$	# false negatives	

ROC curves

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

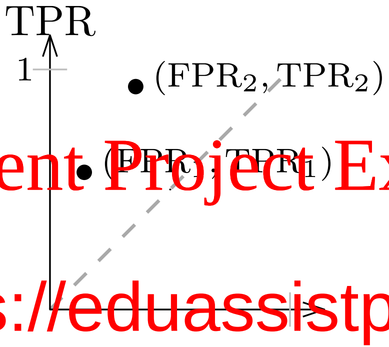

Figure 5: TPR vs FPR plot with two poi

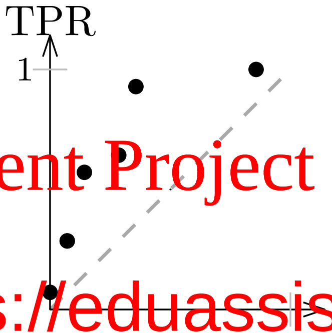

Figure 6: TPR vs FPR plot with many po

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 $[K] := \{1, \ldots, K\}$, determined probability vector $\theta = (\theta_1, \ldots, \theta_K)$

$$K$$

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

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 $\eta(x) = (\eta(x)_1, \ldots, \eta(x)_K)$
  - Conditional probability function: $\eta(x)_k := \Pr(Y = k \mid X = x)$

$/ \quad X)_k]$

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?
- What if we want to eventually use on digits wri Alice and Bob?