IML Exam Grading principles

Johannes Verwijnen

17 December 2014

1 Explaining terms
12 points for explaining terms, 2 points each

(a) Interval scale
An attribute has an interval scale if the differences between values are meaningful, thus operations like addition and subtraction, values like mean and standard deviation, make sense. (2p) Lecture 2, slide 44.

Multiplication and division are not meaningful for interval scale attributes, but ratio scale.

(b) Precision and Recall
In binary classification precision is the ratio of true positive predictions from all positive predictions \( \frac{TP}{TP + FP} \). Recall is the ratio of true positive predictions from all positive classes \( \frac{TP}{TP + FN} \). (1p each) Lecture 3, slide 88.

(c) Gini index
The Gini index is an impurity measure, used to calculate the “goodness” of a split calculated as

\[
Gini(t) = 1 - \sum_{i=0}^{K-1} p(i \mid t)^2
\]

(2p) Lecture 7, slide 164.

(d) One-against-one multiclass classification
This is used when using a linear classifier with a multiclass problem. In this method one creates a linear classifier for each pair of classes using only the data with true class in the pair. When predicting, data gets classified by all these classifiers and the class that gets most “votes” in the pairoffs is the final prediction. (2p) Lecture 8, slide 217.

This method is computationally much more demanding than the other presented, one-against-rest or 1-vs-all classifier.
(e) Generative and discriminative classification

In generative classification the probability of the data point itself \( p(x \mid y) \) is computed, as in the Naïve Bayes classifier, whereas discriminative classification only models \( p(y \mid x) \), or even \( \hat{y} = f(x) \). (1p each) Lecture 5, slide 131.

(f) Ridge regression

Ridge regression is one way to regularize overfitting in linear regression by adding a penalty for complicated solutions (weights with a large norm). It uses the objective function

\[
E_{\text{ridge}} = \sum_{i=1}^{N} (y_i - w^T x_i)^2 + \lambda \|w\|_2^2
\]

with the solution becoming

\[
w^* = (X^T X + \lambda I)^{-1} X^T y
\]

Lecture 8, slide 209.

2 Bayes optimal prediction

We can find the wanted \( p(Y = i \mid x) \) using Bayes’ theorem, thus

\[
p(Y = i \mid x) = \frac{p(x \mid Y = i) p(Y = i)}{p(x)}
\]

The likelihood \( p(x \mid Y = i) \) we have to calculate based on what we know of the labels. The probability densities are thus

\[
p(x \mid Y = 1) = \begin{cases} \frac{1}{5-0} = \frac{1}{5} & \text{if } x \in [0, 5] \\ 0 & \text{otherwise} \end{cases}
\]

\[
p(x \mid Y = 2) = \begin{cases} \frac{1}{8-5} = \frac{1}{3} & \text{if } x \in [5, 8] \\ 0 & \text{otherwise} \end{cases}
\]

\[
p(x \mid Y = 3) = \begin{cases} \frac{1}{6-4} = \frac{1}{2} & \text{if } x \in [4, 6] \\ 0 & \text{otherwise} \end{cases}
\]

(1p)

We can split the total range of \( x \) into subranges based on the distribution of the labels. The marginal probabilities of \( p(Y = i) \) were given in the assignment. We can find the needed \( p(x) \) by the Law of Total Probability

\[
p(x) = \sum_{i=1}^{3} p(x \mid Y = i)
\]
\[ p(x \in [0, 4)) = \frac{1}{2} \times \frac{4 - 0}{9} + 0 + 0 = \frac{2}{9} \]
\[ p(x \in [4, 5)) = \frac{1}{2} \times \frac{5 - 4}{9} + 0 + \frac{1}{6} \times \frac{5 - 4}{2} = \frac{5}{36} \]
\[ p(x \in [5, 6]) = \frac{1}{2} \times \frac{6 - 5}{9} + \frac{1}{3} \times \frac{6 - 5}{3} + \frac{1}{6} \times \frac{6 - 5}{2} = \frac{1}{4} \]
\[ p(x \in (6, 8]) = \frac{1}{2} \times \frac{8 - 6}{9} + \frac{1}{3} \times \frac{8 - 6}{3} + 0 = \frac{1}{3} \]
\[ p(x \in (8, 9]) = \frac{1}{2} \times \frac{9 - 8}{9} + 0 + 0 = \frac{1}{18} \]

(1p)

Now we have all we need to calculate the posterior class probabilities, resulting in: (6p, these were explicitly asked for in the question)

\[ p(Y = 1 | x) \]
\[ p(Y = 2 | x) \]
\[ p(Y = 3 | x) \]

\[
\begin{array}{c|c|c|c}
  x & \frac{p(Y = 1 | x)}{p(Y = 2 | x)} & \frac{p(Y = 2 | x)}{p(Y = 3 | x)} & \frac{p(Y = 3 | x)}{p(Y = 3 | x)} \\
  \hline
  x \in [0, 4) & \frac{\frac{2}{9}}{\frac{5}{36}} = 1 & 0 & 0 \\
  x \in [4, 5) & \frac{\frac{5}{36}}{\frac{1}{3}} = \frac{2}{5} & 0 & \frac{\frac{5}{36}}{\frac{1}{3}} = \frac{3}{5} \\
  x \in [5, 6) & \frac{\frac{1}{2}}{\frac{1}{3}} = \frac{2}{5} & \frac{\frac{1}{2}}{\frac{1}{3}} = \frac{4}{5} & \frac{\frac{1}{2}}{\frac{1}{3}} = \frac{1}{3} \\
  x \in (6, 8] & \frac{\frac{1}{3}}{\frac{1}{3}} = \frac{1}{3} & \frac{\frac{1}{3}}{\frac{1}{3}} = \frac{2}{3} & 0 \\
  x \in (8, 9] & \frac{\frac{1}{18}}{\frac{1}{18}} = 1 & 0 & 0 \\
\end{array}
\]

The Bayes optimal classifier is

\[ f_*(x) = \arg \max_y P(y | x) \]

Thus we get \( Y = 1 \) when \( x \in [0, 4] \lor x \in (8, 9] \), \( Y = 2 \) when \( x \in [5, 8] \) and \( Y = 3 \) when \( x \in [4, 5) \). (2p)

The Bayes error (or risk) is the amount of wrong predictions made, thus we sum the following:

\[ p(Y \neq 1 | x \in [0, 4))p(x \in [0, 4]) = 0 \]
\[ p(Y \neq 3 | x \in [4, 5))p(x \in [4, 5]) = \frac{2}{5} \times \frac{5}{36} = \frac{1}{18} \]
\[ p(Y \neq 2 | x \in [5, 6])p(x \in [5, 6]) = \frac{5}{9} \times \frac{1}{4} = \frac{5}{36} \]
\[ p(Y \neq 2 | x \in (6, 8])p(x \in (6, 8]) = \frac{1}{3} \times \frac{1}{3} = \frac{1}{9} \]
\[ p(Y \neq 1 | x \in (8, 9])p(x \in (8, 9]) = 0 \]

and get a Bayes error of \( \frac{11}{36} \). (2p)

Many used discrete integers \( (x = 0 \ldots 9) \) for the calculations, resulting in wrong answers. If calculations were correct given the circumstances, only 2 points were deducted for that mistake.

3 Overfitting and underfitting

- overfitting definition (training data vs. test)
- underfitting definition (model not sufficiently complex)
- 1NN → each point is itself neighbor, perfect fit!
- in kNN, the larger ks smooth out local differences
- Hastie et al, 2009 -like picture
- high complexity (overfitting) → low bias/high variance, low complexity (underfitting) → high bias/low variance
- validation and cross-validation or complexity penalization as antidotes
- example in linear regression or decision trees

(2p each, minuses for incorrect or irrelevant points)

4 Logistic regression

(2p each subpart, based on completeness, minuspoints for incorrect or irrelevant points)

4.1 what type of learning problems can it be used for
Binary classification

4.2 any particular assumptions about the inputs
Training input \((x, y)\), where \(y \in \{1, 0\}\) and \(x \in \mathbb{R}^n\).

4.3 what kind of output is produced
Probabilities in \([0, 1]\)

4.4 precise mathematical formulation
\[
P(y = +1 \mid x) = \frac{1}{1 + \exp(-w^T x)}
\]

4.5 parameters, important generalizations, or other variants
Regularization possible by adding \(\lambda \|w\|_2^2\)
4.6 comparison to one or two other algorithms that can be used for similar problems
Perceptron, support vector machine

5 Hierarchical clustering

(a) basics
- Tree-like structure, similar data close to each other (2p)
- Divisive or agglomerative approach (2p)
- Input: data vectors \( \mathbf{x}_i, i = 1 \ldots N \) (2p)
- Output: Cluster assignments and dissimilarities (dendogram) (2p)

(b) single link vs complete link
- single link: minimum dissimilarity between data points in different clusters (1p)
- complete link: maximum dissimilarity between data points in different clusters (1p)
- example (2p)