Weekly exercises

IN4310 / IN3310 — Deep Learning for Image Analysis

Department of Informatics, University of Oslo

Performance estimation

1 Average precision in extreme cases

a

Suppose the set used for evaluation consists of 11 samples, of which 3 are actual positives, and those 3 have *higher* prediction score for the positive class than the other 8 samples. What is the average precision?

$$AP = 1$$

b

Suppose the set used for evaluation consists of 11 samples, of which 3 are actual positives, and those 3 have *lower* prediction score for the positive class than the other 8 samples. What is the average precision?

$$AP = (1/9 + 2/10 + 3/11)/3 \approx 0.1946$$

c

Suppose the set used for evaluation consists of N samples, of which R are actual positives, and those R have *lower* prediction score for the positive class than

the other $N - R$ samples. What is the average precision?

Hint: The answer can be an expression that contains a summation and depends on R and N. You will (likely) not be able to write it without a summation.

$$AP = \frac{1}{R} \sum_{r=1}^{R} \frac{r}{N - R + r}$$

2 Average precision of random predictor

Suppose the set used for evaluation consists of N samples, of which R are actual positives. A predictor is used to calculate prediction scores for the positive class for each sample.

a

Assume the ranking of the scores evenly distributes the R actual positive samples. More specifically, after sorting the samples by their scores in descending order and letting 1 be the first index (the sample with highest score), assume that the first actual positive sample has index $\frac{N}{R}$ (and that this is an integer), the second actual positive sample has index $2\frac{N}{R}$, and so on, making the l-th actual positive sample appear at index $l\frac{N}{R}$ and the last actual positive sample appear at index $R\frac{N}{R} = N$. What is the average precision?

Hint 1: Calculate the precision at k ($P@k$) for every index of an actual positive sample, i.e. for $k = l\frac{N}{R}$ with $l = 1, 2, \ldots, R$.

Hint 2: It is possible to write the average precision without any summation.

After sorting the scores, the precision at k is defined as:

$$P@k = \frac{1}{k} \sum_{i=1}^{k} y_i$$

To calculate AP, we need to calculate $P@k$ at every index of an actual positive sample, which is at $k = l\frac{N}{R}$ for $l = 1, 2, \ldots, R$. Notice that there are l actual positive samples among the k first samples when $k = l\frac{N}{R}$ for $l = 1, 2, \ldots, R$. Therefore, inserting $l\frac{R}{N}$ for k in the definition of $P@k$ yields:

$$P@(l\frac{R}{N}) = \frac{l}{l\frac{N}{R}} = \frac{R}{N}$$

This is valid for $k = l\frac{N}{R}$ with $l = 1, 2, \ldots, R$, which is precisely the non-zero terms in the summation used to calculate the average precision:

$$AP = \frac{1}{R}\sum_{k=1}^{N} y_k P@k = \frac{1}{R}\sum_{l=1}^{R} P@(l\frac{R}{N}) = \frac{1}{R}\sum_{l=1}^{R}\frac{R}{N} = \frac{R}{N}$$

The AP for this predictor, which can be seen as a random classifier, is thus equal to the proportion of actual positives in the evaluation set.

b

Again assume that the ranking of the scores evenly distributes the R actual positive samples, but this time the order is shifted such that the first actual positive sample has index 1. More specifically, after sorting the samples by their scores in descending order and letting 1 be the first index (the sample with highest score), the first actual positive sample has index 1, the second actual positive sample has index $1 + \frac{N}{R}$ (and this is assumed to be an integer), and so on, making the l-th actual positive sample appear at index $1 + (l-1)\frac{N}{R}$ and the last actual positive sample appear at index $1 + (R-1)\frac{N}{R}$. What is the average precision?

Hint: The answer can be an expression that contains a summation and depends on R and N. You will (likely) not be able to write it without a summation.

This time, the indices of the actual positive samples are $1 + (l-1)\frac{N}{R}$ for $l = 1, 2, \ldots, R$, and for such k there are l actual positive samples among the k first samples. Therefore, inserting $1 + (l-1)\frac{N}{R}$ for k in the definition of $P@k$ yields:

$$P@(1 + (l-1)\frac{N}{R}) = \frac{l}{1 + (l-1)\frac{N}{R}}$$

This is valid for $k = 1 + (l-1)\frac{N}{R}$ with $l = 1, 2, \ldots, R$, which is precisely the non-zero terms in the summation used to calculate the average precision:

$$\begin{aligned}
AP &= \frac{1}{R}\sum_{k=1}^{N} y_k P@k \\
&= \frac{1}{R}\sum_{l=1}^{R} P@(1 + (l-1)\frac{R}{N}) \\
&= \frac{1}{R}\sum_{l=1}^{R}\frac{l}{1 + (l-1)\frac{N}{R}} \\
&= \frac{1}{R}\sum_{l=1}^{R}\frac{l}{1 - \frac{N}{R} + l\frac{N}{R}}
\end{aligned}$$

c

Will the average precision in subtask **b** be lower or higher than the average precision in subtask **a**?

Intuitively, the average precision in subtask **b** will be higher because the sample with highest score is an actual positive and the actual positive samples are otherwise evenly distributed in both cases. We can show this concretely by proving that each of the R terms in the summation is higher for the predictor in subtask **b** than for the predictor in subtask **a**, i.e. that:

$$\frac{l}{1-\frac{N}{R}+l\frac{N}{R}} > \frac{R}{N}$$

This is indeed the cases if $N > R$, which simply means that not all samples are positives, because then $\frac{N}{R} > 1$ (yes, strictly speaking with the additional assumption that $R \neq 0$, but that simply means that there are at least 1 actual positive sample), which again implies that $1-\frac{N}{R} < 0$. Because adding a negative value to a denominator makes the fraction larger, we have that:

$$\frac{l}{1-\frac{N}{R}+l\frac{N}{R}} > \frac{l}{l\frac{N}{R}} = \frac{R}{N}$$

This is valid if $N > R$. If $N = R$, then the two average precisions are identical; in fact, both would then be 1 as all samples are actual positives in this case.

3 Average precision and accuracy of a linear classifier

Consider a linear classifier $s(x) = wx + b$ classifying samples as positives if $s(x) > 0$ and otherwise as negatives. w and b are trainable parameters.

a

Which of the trainable parameters will impact the resulting average precision calculated using the prediction scores for the positive class, $s(x)$? Please explain.

w determines the direction in which $s(x)$ increases. Different w can result in very different ranking of the samples by their $s(x)$. As a simple example, consider a set with three points $x_1 = (1, 0)$, $x_2 = (1, 1)$, and $x_3 = (0, 2)$. If $w = (1, 0)$, then $wx_1 = wx_2 = 1$ and $wx_3 = 0$, so x_1 and x_2 is ranked first and x_3 is ranked last. If $w = (0, 1)$, then $wx_1 = 0$, $wx_2 = 1$, and $wx_3 = 2$, so x_3 is ranked first, x_2 second, and x_1 last. Because a different ranking of the samples can impact the average precision, w will impact the average precision.

b shifts the decision boundary along the direction defined by w but does not influence the ranking of the samples by $s(x)$. Therefore, b will not influence the average precision.

b

Which of the trainable parameters will impact the resulting accuracy? Please explain.

For the accuracy, also b will have an impact because it will define the threshold at which wx should be considered large enough to be classified as representing a positive sample. Different thresholds result in different sets of samples being classified as positives and negatives, which will impact the accuracy.

4 Area under the receiver operating characteristic curve (AUROC)

a

Assume that we observe that a model obtains an AUROC of 1 when using a particular evaluation set. What does that imply for the ranking of the prediction scores for the positive class in that evaluation set?

It implies that all scores for actual positive samples are greater than all scores for actual negative samples. Another way to formulate this is that there exists a threshold such that all samples with a score below the threshold are actual negatives and all samples with a score above the threshold are actual positives.

Note that the highest score for an actual negative has to be strictly less than the lowest score for an actual positive. If this was not the case, i.e. if there existed a negative sample with a score greater than or equal to the score of a positive sample, then it would not be possible to place a threshold such that the sensitivity and specificity of the resulting classifier are both 1, and consequently the AUROC would have been less than 1.

It is also possible to see this by considering the pairs of scores where one is from an actual negative sample and the other is from an actual positive sample. If the AUROC is 1, then the score for the actual positive sample has to be greater than the score for the actual negative sample because that is the only way it will be fully counted in the nominator of the fraction and the denominator is the number of all possible pairs.

b

Suppose the set used for evaluation consists of N samples, of which 2 are actual positives. One of these 2 actual positives has a higher prediction score for the positive class than the other $N-1$ samples. The other of these 2 actual positives has a prediction score for the positive class that is greater than 80% of the prediction scores for the positive class for actual negative samples (and there are no ties, so 20% of the actual negative samples have a higher score than 1 of the 2 actual positives). What is the AUROC?

Let us consider the pairs of scores where one is from an actual negative sample and the other is from an actual positive sample. For 1 of the 2 actual positives, the score of the actual negative sample is less for all pairs. For the other of the 2 actual positives, the score of the actual negative sample is less for 80% of the pairs and greater for 20% of the pairs. Because the number of pairs is identical for each of the 2 actual positives, this implies that for $\frac{100\%+80\%}{2} = 90\%$ of all pairs the score of the actual negative sample is less than the score of the actual positive sample and the opposite is the case for the remaining 10% of the pairs. Thus, the $AUROC = 0.9$.

We could alternatively have plotted the receiver operating characteristic (ROC) curve and calculated the area under this curve. The ROC curve increases from sensitivity 0 to 0.5 for 1−specificity 0, then increases from 1−specificity 0 to 0.2 for sensitivity 0.5, then increases from sensitivity 0.5 to 1 for 1−specificity 0.2, and finally increases from 1−specificity 0.2 to 1 for sensitivity 1. The area under this curve is therefore:

$$AUROC = 0.2 * 0.5 + 0.8 * 1 = 0.1 + 0.8 = 0.9$$

c

Suppose the set used for evaluation consists of N samples, of which 4 are actual positives. One of these 4 actual positives has a higher prediction score for the positive class than the other $N-1$ samples. Two of the other of these 4 actual positives has a prediction score for the positive class that is greater than 80% of the prediction scores for the positive class for actual negative samples (and there are no ties, so 20% of the actual negative samples have a higher score than these 2 of the 4 actual positives). The last of these 4 actual positives has a prediction score for the positive class that is greater than 60% of the prediction scores for the positive class for actual negative samples (and there are no ties, so 40% of the actual negative samples have a higher score than 1 of the 4 actual positives). What is the AUROC?

Let us again consider the pairs. For 1 of the 4 actual positives, the score of the actual negative sample is less for all pairs. For 2 of the other actual positives,

the score of the actual negative sample is less for 80% of the pairs and greater for 20% of the pairs. For the last of the 4 actual positives, the score of the actual negative sample is less for 60% of the pairs and greater for 40% of the pairs. Because the number of pairs is identical for each of the 4 actual positives, this implies that for $\frac{100\%+2*80\%+60\%}{4} = 80\%$ of all pairs the score of the actual negative sample is less than the score of the actual positive sample and the opposite is the case for the remaining 20% of the pairs. Thus, the $AUROC = 0.8$.

It is of course also here possible to plot the ROC curve and calculated the area under this curve in order to obtain the same result. The ROC curve increases from sensitivity 0 to 0.25 for 1−specificity 0, then increases from 1−specificity 0 to 0.2 for sensitivity 0.25, then increases from sensitivity 0.25 to 0.75 for 1−specificity 0.2, then increases from 1−specificity 0.2 to 0.4 for sensitivity 0.75, then increases from sensitivity 0.75 to 1 for 1−specificity 0.4, and finally increases from 1−specificity 0.4 to 1 for sensitivity 1. The area under this curve is therefore:

$$AUROC = 0.2 * 0.25 + 0.2 * 0.75 + 0.6 * 1 = 0.2 + 0.6 = 0.8$$

d

Suppose the evaluation set and scores are the same as described in subtask **c**. In order to obtain a dichotomous classifier, the score is thresholded such that 3 of the 4 actual positives and 20% of the actual negatives are classified as positives, while the last actual positive and 80% of the actual negatives are classified as negatives. What is the AUROC of this binary classifier?

There are again multiple approaches which will all lead to the same result. The easiest is to remember that the AUROC of a binary classifier is identical to the balanced accuracy of the classifier. Because the sensitivity of the classifier is 75% and the specificity of the classifier is 80%, the balanced accuracy and thereby also the AUROC is 0.775 in this case.