Solutions to Test 2 for Uncertainty in AI class, Fall 2025

1. Suppose that we know the measurement results $\widetilde{x}_1, \ldots, \widetilde{x}_n$, the data processing algorithm $y = f(x_1, \ldots, x_n)$, and the standard deviations $\sigma_1, \ldots, \sigma_n$. Describe how to best estimate the standard deviation σ of the data processing result y when n is small (provide the formula) and how to estimate it when n is large (just explain the idea).

Answer: When n is small, we can use an analytical formula

$$\sigma = \sqrt{c_1^2 \cdot \sigma_1^2 + \ldots + c_n^2 \cdot \sigma_n^2},$$

where

$$c_i \stackrel{\text{def}}{=} \frac{\partial f}{\partial x_i}.$$

When n is large, it is more effective to use Monte-Carlo method, when we simulate the measurement errors and then use the simulated values Δy to estimate σ.

2. Suppose that we know the measurement results $\widetilde{x}_1, \ldots, \widetilde{x}_n$, the data processing algorithm $y = f(x_1, \ldots, x_n)$, and the upper bounds $\Delta_1, \ldots, \Delta_n$ on the absolute values of the measurement errors. Describe how to best estimate the upper bound Δ of the absolute value of the approximation error of the measurement result y when n is small (provide the formula) and how to estimate it when n is large (just explain the idea).

Answer: When n is small, we can use an analytical formula

$$\Delta = |c_1| \cdot \Delta_1 + \ldots + |c_n| \cdot \Delta_n,$$

where

$$c_i \stackrel{\text{def}}{=} \frac{\partial f}{\partial x_i}.$$

When n is large, it is more effective to use Monte-Carlo method that uses Cauchy distribution, when we simulate the measurement errors and then use the simulated values Δy to estimate Δ.

3. Suppose that we use bisection to compute $\sqrt{2}$, i.e., the solution to the equation $f(x) = 0$ when $f(x) = x^2 - 2$. We know that $f(1) = 1^2 - 2 = 1 - 2 = -1 < 0$ and that $f(3) = 3^2 - 2 = 9 - 2 = 7 > 0$, so we know that the solution is somewhere on the interval $[1, 3]$. When we follow bisetion method, what would we do next, and what will be the resulting new narrower interval?

Answer. We take the midpoint $m = (1 + 3)/2 = 2$ of the given interval, and compute $f(m) = f(2) = 2^2 - 1 = 4 = 1 = 3 > 0$. Since $f(1) < 0$ and $f(2) > 0$, we know the function $f(x)$ changes sign on the interval $[1, 2]$, so it must attain the value 0 somewhere on this interval.

4. Suppose that $y = f(x_1, x_2) = x_1^2 - x_2^2$. Suppose that with confidence 0.5, experts believe that the actual value of x_1 is in the interval $[1, 2]$, and that actual value of x_2 is in the interval $[2, 3]$. Describe the corresponding alpha-cut for y.

Hint: The above function $y = f(x_1, x_2)$ is strictly increasing with respect to x_i and strictly decreasing with respect to x_2 – when both x_i are positive. So, when the inputs x_i are located in intervals, this function:

- attains its smallest value when x_1 is the smallest possible and x_2 is the largest possible, and
- attains its largest value when x_1 is the largest possible and x_2 is the smallest possible.

Answer. The alpha-cut is:

$$[1^2 - 3^2, 2^2 - 2^2] = [1 - 9, 4 - 4] = [-8, 0].$$

5. Suppose that our degree of confidence in a statement A is 0.8, in a statement B is 0.7, and in a statement C is 0.6. Suppose that we use min as "and" and max as "or". What is our estimate for the degree of confidence in a composite statement $A \vee (\neg B \,\&\, C)$?

Answer. In general, the desired degree is equal to $d = f_\vee(a, f_\&(f_\neg(b), c))$, where a b, and c are our degrees of confidence in statements A, B, and C. For $f_\neg(x) = 1 - x$ and for our choice of "and"- and "or"-operations, we have $d = \max(a, \min(1 - b, c))$. For given degrees of confidence a, b, and c, we get

$$d = \max(0.8, \min(1 - 0.7, 0.6)) = \max(0.8, \min(0.3, 0.6)) = \max(0.8, 0.3) = 0.8.$$

6. If we have two alternatives, with gains $[2, 3]$ and $[1, 4]$, which of them are possibly optimal? definitely optimal? Which of the alternatives should we choose if Hurwicz coefficient α_H is 0.4?

Answer. An alternative is definitely optimal if its lower endpoint is larger than or equal to all other upper endpoints. One can check that in this case, there is no such alternative:

- for Alternative 1, its lower endpoint 2 is not larger than the upper endpoint 4 of the second alternative;
- for Alternative 2, its lower endpoint 1 is not larger than the upper endpoint 3 of the first alternative.

An alternative is possibly optimal if its upper endpoint is larger than or equal to all the lower endpoints – i.e., equivalently, to the maximum of the lower endpoints. In our case, this maximum is equal to $\max(1, 2) = 2$. So, an alternative is possible optimal if its upper endpoint is larger than or equal to 2. This property is true for both Alternatives 1 and 2, so they are both possible optimal.

In the Hurwicz approach, we replace each interval $[\underline{x}, \overline{x}]$ with the value $x = \alpha_H \cdot \overline{x} + (1 - \alpha_H) \cdot \underline{x}$, and select the alternative for which this number if the largest. So, for $\alpha_H = 0.2$, we have:

$$x_1 = 0.4 \cdot 3 + (1 - 0.4) \cdot 2 = 1.2 + 0.6 \cdot 2 = 1.2 + 1.2 = 2.4,$$

$$x_2 = 0.4 \cdot 4 + (1 - 0.4) \cdot 1 = 1.6 + 0.6 \cdot 1 = 1.6 + 0.6 = 2.2.$$

so we select Alternative 1.

7. Prove that for each final invariant optimality criterion, the optimal alternative is itself invariant.

Answer. An optimality criterion is a pair of relations $(>, \sim)$, where $a > b$ means that a is better than b, and $\sim$ means that a and b are of the same quality. An optimality criterion is final if there is exactly one optimal alternative. The criterion is invariant with respect to some transformation T if $a > b$ implies $T(a) > T(b)$ and $a \sim b$ implies $T(a) \sim T(b)$. Let $a_{\rm opt}$ be an optimal alternative, this means that for every alternative a we have either $a_{\rm opt} > a$ or $a_{\rm opt} \sim a$. In particular, for every a, we have either $a_{\rm opt} > T^{-1}(a)$ or $a_{\rm opt} \sim T^{-1}(a)$. Since the optimality criterion is invariant, this implies that either $T(a_{\rm opt}) > T(T^{-1}(a)) = a$ or $T(a_{\rm opt}) \sim a$. This is true for all a. By definition of an optimal alternative, this means that $T(a_{\rm opt})$ is optimal. Since the optimality criterion is final, there is only one optimal alternative, so indeed $T(a_{\rm opt}) = a_{\rm opt}$.

8. In the classical logic, the equivalence $A \equiv B$ is defined as $f_\equiv(0,0) = f_\equiv(1,1) = 1$ and $f_\equiv(0,1) = f_\equiv(1,0) = 0$. Use linear interpolation to come up with a formula for $f_\equiv(a,b)$, and use this formula to come up with the fuzzy value $f_\equiv(0.7,0.8)$. Why do we use linear interpolation?

Answer. The general formula for linear interpolation has the following form: if we know the values $y_1 = f(x_1)$ and $y_2 = f(x_2)$, then for every other x, we have

$$f(x) = y_1 + \frac{y_2 - y_1}{x_2 - x_1} \cdot (x - x_1).$$

Let us first apply linear interpolation to find the values $f_\equiv(0,x)$ for different x. For $x_1 = 0$, we have $y_1 = f_\equiv(0,0) = 1$, while for $x_2 = 1$, we have $y_2 = f_\equiv(0,1) = 0$. Thus, we have

$$f_\equiv(0,x) = 1 + \frac{0-1}{1-0} \cdot (x-0) = 1 - x.$$

Let us now apply linear interpolation to find the values $f_\equiv(1,x)$ for different x. For $x_1 = 0$, we have $y_1 = f_\equiv(1,0) = 0$, while for $x_2 = 1$, we have $y_2 = f_\equiv(1,1) = 1$. Thus, we have

$$f_\equiv(1,x) = 0 + \frac{1-0}{1-0} \cdot (x-0) = x.$$

Finally, let us apply linear interpolation to find the value $f_\equiv(x,b)$. For $x_1 = 0$, we have $y_1 = f_\equiv(0,b) = 1-b$, while for $x_2 = 1$, we have $y_2 = f_\equiv(1,b) = b$. Thus, we have

$$f_\equiv(x,b) = 1 - b\frac{b-(1-b)}{1-0} \cdot (x-0) = 1 - b + (2b-1)\cdot x = 1 - b + 2b \cdot x - x.$$

For $x = 0.7$ and $b = 0.8$, we get

$$f_\equiv(0.7,0.8) = 1 - 0.7 + 2 \cdot 0.7 \cdot 0.8 - 0.8 = 0.3 + 2 \cdot 0.56 - 0.8 = 0.3 + 1.12 - 0.8 =$$

$$1.42 - 0.8 = 0.62.$$

We use linear interpolation because it is the only one that is invariant with respect to change of measuring unit and starting point and is, therefore optimal with respect to any invariant optimality criterion.

9. Explain, in all necessary details, how to use invariance to explain why ReLU is the optimal activation function. You can use the result that optimal alternative is invariant.

Answer. We want to select an activation function $s(x)$ that is optimal, and it is reasonable to require that the optimality criterion should be invariant with respect to the change of a measuring unit. Thus, the optimal activation function should be thus invariant. In other words, we want to make sure that if $y = s(x)$ and we select a new measuring unit, i.e., switch to new numerical values $x' = \lambda \cdot x$ and $y' = \lambda \cdot y$, then for these new values x' and y', we will have the exact same dependence:

$$y' = s(x').$$

Substituting the expressions $x' = \lambda \cdot x$ and $y' = \lambda \cdot y$ into this formula, we conclude that $\lambda \cdot y = s(\lambda \cdot x)$. Here, $y = s(x)$, so we conclude that $s(\lambda \cdot x) = \lambda \cdot s(x)$ for all possible x and $\lambda > 0$.

For $x = 1$, we conclude that $s(\lambda) = \lambda \cdot s(1)$. If we denote $s(1)$ by c_+, and rename λ into z, we conclude that for all $z > 0$, we get

$$s(z) = c_+ \cdot z.$$

For $x = -1$, we conclude that $s(-\lambda) = \lambda \cdot s(-1)$. If we denote $-s(-1)$ by c_- (so that $s(-1) = -c_-$) and denote $-\lambda$ by z (so that $\lambda = -z$), we conclude that for all negative values z, we have

$$s(z) = (-c_-) \cdot (-z) = c_- \cdot z.$$

Thus, we conclude that the activation function $s(z)$ should have the following *piecewise linear* form (see, e.g., [**?**]):

- for $z > 0$, we have $s(z) = c_+ \cdot z$;
- for $z < 0$, we have $s(z) = c_- \cdot z$.

In particular, for $c_- = 0$ and $c_+ = 1$, we get ReLU. One can show that because of the linear layers, the use of any other such piecewise linear activation function is equivalent to using ReLU.