Uczenie Maszynowe, egzamin 0

Egzamin składa się z 20 pytań wielokrotnego wyboru, każde jest warte 1 punkt. Pytania mają różną liczbę odpowiedzi N, gdzie prawidłowe jest P odpowiedzi. Prawidłowych odpowiedzi jest między 1 a N (włącznie). Każda prawidłowa odpowiedź to $+1/P$ punktów, nieprawidłowa $-1/P$, ale nie można otrzymać mniej niż 0 punktów za pytanie. Można zaznaczyć między 1 a N odpowiedzi.

Przykładowo, jeżeli pytanie ma 5 odpowiedzi i 3 prawidłowe, to zaznaczenie 2 prawidłowych i 1 nieprawidłowej daje $+2/3 - 1/3 = 1/3$.

Zadanie 1

Wiemy, że w N-wymiarowej przestrzeni cech mamy M próbek treningowych ($N > 100, M > 100$) i K próbek testowych. Zakładamy, że $M = N + 1$, a obie miary są dwie klasy. Dane tych klas pochodzą z dwóch przesuniętych względem siebie rozkładów normalnych (ang. *normal/Gaussian*) o jednakowych macierzach kowariancji. Zbiory testowe i treningowe są zrównoważone (ang. *balanced*). Próbki treningowe mają etykiety przynależności do poszczególnych klas (*class label*) i nie leżą na rozmaitości (ang. *manifold*) o mniejszym rozmiarz niż N. Zaznacz prawidłowe odpowiedzi. Przez dokładność rozumiemy (ang. *accuracy*).

a) Żaden z pozostałych odpowiedzi nie jest prawdziwa. Wynik klasyfikacji zależy od wielu czynników, w tym od rozkładu danych, separowalności klas, iczby i wyboru próbek w zbiorach treningowych i testowych oraz wybranych metod przetwarzania i klasyfikacji.

b) Zastosowanie klasyfikatora liniowego do nazw tym przypadku zawsze wynik klasyfikacji danych treningowych z dokładnością $1.0\,(100\%)$, bez względu na to, z jakich dokładnie rozkładów pochodzą próbki.

c) Skonstruowanie klasyfikatora liniowego danych dla danych treningowych da nam zawsze wynik klasyfikacji danych testowych o dokładności $\geq 0.51\,(51\%)$, bo dwa przesunięte względem siebie rozkłady Gaussa o jednakowych macierzach kowariancji są zawsze sparowalne liniowo ang. *linearly separable*)

d) Gdy M i N zmierzają do nieskończoności, dla $M \leq 2N+1$ z prawdopodobieństwem równym $1.0\,(100\%)$ można znaleźć klasyfikator liniowy separujący z dokładnością $1.0\,(100\%)$ dane treningowe na dwie zdefiniowane klasy.

e) Zastosowanie klasyfikatora liniowego da nam zawsze wyniki klasyfikacji dla danych testowych o dokładności $\geq 0.51\,(51\%)$, ale po transformacji PCA do mniejszego wymiaru $L < N/2$.

Zadanie 2

Mamy zbiór danych do klasyfikacji i 3 cech A, B, C. Każda cecha wnosi do klasyfikatora 0.25 dokładności (ang. *accuracy*), natomiast całkowita dokładność klasyfikatora to 0.8. Problem z cechą B polega na tym, że występuje w parze z inną cechą obniża jakoś

klasyfikatora o 0.1. Stosując podejście Shapley'a, wyznacz wartość Shapley'a dla poszczególnych cech i ustaw cechy w kolejności od najważniejszej do najmniej ważnej. Zaznacz prawidłowe odpowiedzi.

a) $C = 0.5$

b) $A = C, B$

c) $A = 0.4$

d) $B = 0.15$

e) A, B, C

Zadanie 3

Wybierz warunki, które musi spełniać destylacja wiedzy z modelu-nauczyciela (ang. *teacher model*) do modelu-studenta (ang. *student model*). Zakładamy, że pojemność informacyjna (ilość parametrów, ang. *capacity*) studenta jest mniejsza niż nauczyciela.

a) Wymiar wektora logitów (i wyników softmaxu) w obu sieciach neuronowych musi być taki sam.

b) Nauczyciel i student muszą być modelami reprezentowanymi przez sieci neuronowe.

c) Student nie może być „wygrywającym biletem" (ang. *winning ticket*) dla nauczyciela.

d) Musi być tylko jeden nauczyciel.

e) Nauczyciel może osiągnąć gorszy rezultat (np. *accuracy, loss*) niż student.

Zadanie 4

Zaznacz prawidłowe stwierdzenia dotyczące przycinania wag sieci neuronowych.

a) Przycinanie wag zawsze prowadzi do degradacji jakości sieci (biorąc pod uwagę np. *accuracy, loss*).

b) Przycinanie wag sieci i ponowne przywracanie niektórych połączeń (ang. *pruning with rehabilitation*) z reguły poprawia jakość sieci.

c) Ustawienie niektórych wag na wartość 1 i eliminacja innych, to radykalna metoda kwantyzacji sieci, którą można uczynić, wykorzystując schemat destylacji wiedzy.

d) Załóżmy, że mamy 2 warstwy neuronów: wejściową m neuronów oraz wyjściową k neuronów. Dokonujemy faktoryzacji macierzy wag W, stosując metodę SVD. Wybieramy t największych wartości własnych. W ten sposób możemy zmniejszyć złożoność obliczeniową z $O(mk)$ do $O(mt)$.

e) Wypłaszczona konwolucja (ang. *flattened convolution*) w sieciach konwolucyjnych (CNNs) polega na zastąpieniu kształtu $C \times Y \times X$ (C — liczba kanałów, X i Y — wymiary obrazu) konwolucjami $C \times 1 \times 1$, $1 \times Y \times 1$, $1 \times 1 \times X$. Redukuje to liczbę parametrów, a reguły polepszając wydajność sieci.

Zadanie 5

Chcemy otrzymać wyjaśnienie (interpretację) predykcji nieinterpretowalnego klasyfikatora (ang. *outcome explanation*). Zwrócił on pewną klasę dla obrazu X (zwanego dalej bazowym obrazem). Możemy wykorzystać tutaj procedurę LIME. Pozostaw prawidłową sekwencję kroków dla tej procedury, oznaczając w odpowiedzi redundantne lub błędne kroki:

a) Dokonujemy podziału obrazu przy pomocy siatki na $N \times N$ komórek, gdzie superpikselem jest komórka.

b) Dokonujemy segmentacji obrazu X, tworząc superpiksele.

c) Dokonujemy augmentacji obrazu X, maskując w sposób losowy jego niektóre superpiksele.

d) Dokonujemy augmentacji obrazu X, mieszając w sposób losowy superpiksele.

e) Znajdujemy etykiety stworzonych kopii, wykorzystując oryginalny, nieinterpretowalny model bazujący na danych.

f) Znajdujemy softmax stworzonych kopii, wykorzystując oryginalny, nieinterpretowalny model bazujący na danych.

g) Znajdujemy wagi stworzonych obrazów w zależności od wartości funkcji softmax, odpowiadającej każdej z kopii.

h) Znajdujemy wagi stworzonych obrazów w zależności od kernelowej odległości (najczęściej z kernelem gaussowskim) zaburzonych kopii od orygnialnego obrazka X.

i) Wybieramy pomniejszony, w stosunku do oryginalnego, model (np. *pruning, knowledge distillation*) i wykorzystujemy go do klasyfikacji wygenerowanych kopii bazowego obrazu.

j) Wybieramy liniowy klasyfikator uwzględniający wagi punktów i wykorzystujemy go do klasyfikacji wygenerowanych kopii bazowego obrazu.

k) Wartości bezwzględne współczynników stojących przy odpowiednich cechach określają ich wagę decydującą o wyborze dla obrazka X odpowiedniej klasy.

l) Wartości bezwzględne i znak stojący przy odpowiednich cechach określają ich wpływ i wagę decydującą o wyborze dla obrazka X odpowiedniej klasy.

Zadanie 6

Zaznacz prawidłowe odpowiedzi dotyczące podstawowych własności szeregów czasowych.

a) Większość algorytmów dla szeregów czasowych zakłada czas ciągły (np. minuty) i wartości całkowite (np. liczba sprzedanych danych sztuk).

b) Nieregularnie próbkowane szeregi czasowe (ang. *irregularly sampled time series*) wymagają imputacji, zazwyczaj przez wypełnienie ostatnią znaną wartością.

c) Szereg składający się tylko z trendu i sezonowości, z resztą będącą białym szumem (ang. *white noise*), nie jest prognozowalny.

d) Algorytm dekompozycji STL może zrealizować dekompozycję multiplikatywną przy połączeniu z odpowiednią transformacją.

e) W celu stacjonaryzacji szeregów wykonuje się operację różnicowania (ang. *differencing*), którą można zautomatyzować z pomocą testów statystycznych.

f) Różnicowanie szeregu usuwa trend i/lub sezonowość.

g) Szereg o bardzo wysokiej autokwariancji (ang. *autocovariance*) jest trudniej prognozowalny od szeregów o niskiej autokowariancji.

h) Transformacja Boxa-Coxa gwarantuje homoskedastyczną wariancję oraz rozkład normalny wartości w zbiorze treningowym i testowym, ułatwiając trening i prognozowanie.

Zadanie 7

Zaznacz prawidłowe odpowiedzi dotyczące ewaluacji prognozowania szeregów czasowych.

a) Dla danych finansowych będących błędzeniem losowym (ang. *random walk*), predykcja ostatniej wartości (ang. *predict last / naive*) jest dobrym lub optymalnym *baseline*'m.

b) Należy zawsze wypróbować odpowiedni *baseline*, np. przewidywanie mediany, bo nawet złożone modele neuronowe mogą okazać się od niego gorsze.

c) Strategia tzw. *expanding window* daje realistyczne i precyzyjne oszacowanie wyników, ale jest kosztowna obliczeniowo.

d) Wadą metryk nieznormalizowanych (ang. *scale-dependent*) jak MAE jest brak możliwości łatwego porównania ich wartości dla szeregów o różnej skali wartości. Wadę tę zwalczają metryki znormalizowane (ang. *scale-independent*) jak MASE.

e) Metryki MAPE i SMAPE mają liczne problemy, w szczególności niestabilność numeryczną dla małych wartości i asymetrię, dlatego nie powinno się z nich korzystać lub zachować daleko idącą ostrożność.

f) Błędy modelu w idealnym przypadku powinny mieć rozkład normalny wycentrowany na zerze i nie mieć autokorelacji, co można sprawdzić za pomocą testów statystycznych, np. Anderson-Darling, Ljung-Box.

Zadanie 8

Zaznacz prawidłowe odpowiedzi dotyczące modeli statystycznych do prognozowania szeregów czasowych.

a) Modele ARIMA są modelami z długą pamięcią (ang. *long memory*), bo są oparte o modelowanie autokorelacji AR(p) i korektę błędów MA(q), a pomaga im w tym różnicowanie I(d).

b) Obliczanie modelu SARIMAX w całości redukuje się do jednej, wieloczęściowej regresji liniowej, estymowanej metodą najmniejszych kwadratów (ang. *OLS*).

c) Zarówno modele ETS, jak i ARIMA wymagają preprocessing'u, jak usuwanie trendu (ang. *detrending*) czy sezonowości (ang. *deseasonalizing / seasonal adjustment*), realizowanych typowo przez różnicowanie (ang. *differencing*).

d) Wszystkie modele z rodziny ETS mają swoje odpowiedniki wśród modeli ARIMA, dzięki temu, że obie grupy potrafią łatwo uwzględniać cechy egzogeniczne (ang. *exogenous variables*).

e) Kiedy nie wiemy, czy warto wybrać LightGBM zamiast ETS lub ARIMA, to używamy modelu, który po treningu dał najlepszą wartość kryterium AIC (ang. *Akaike's information criterion*).

f) Dla modeli regresji ML, zastosowanie kodowania cyklicznego (ang. *cyclical encoding*) oraz cech Fouriera (ang. *Fourier features*) daje w wyniku 2 dodatkowe cechy, po jednej na użytą metodę.

g) Dużą zaletą modeli ETS i ARIMA jest łatwość ich automatyzacji algorytmami AutoETS i AutoARIMA, używającymi np. testów statystycznych czy krokowego przeszukiwania siatki hiperparametrów (ang. *stepwise tuning*).

Zadanie 9

Zaznacz prawidłowe odpowiedzi dotyczące modeli neuronowych do prognozowania szeregów czasowych.

a) Istotnym zastosowaniem modeli neuronowych są złożone, długie, często wielowymiarowe szeregi czasowe.

b) Zaletą modeli neuronowych jest możliwość prognozowania wielu kroków wprzód (ang. *direct multi-step / DMS forecasting*), w odróżnieniu od modeli statystycznych, typowo prognozujących autoregresyjnie (ang. *autoregressive / iterative multi-step forecasting*) w każdym kroku.

c) Dla szeregów wielowymiarowych o mocnych zależnościach (ang. *cross-series correlation*) szczególnie dobre będą modele liniowe, jak np. NLinear, ze względu na odporność na szum.

d) Zaletą architektury N-BEATS jest jej interpretowalność, ale w klasycznej wersji działa ona tylko dla jednowymiarowych szeregów (ang. *univariate*).

e) TSMixer osobno uczy się relacji wewnątrz szeregu (w czasie) i pomiędzy szeregami (cechami), co daje dużą efektywność, a jednocześnie regularizuje model w porównaniu do modelu, który porównywałby wszystkie kombinacje naraz.

f) Większość transformerów dla szeregów czasowych to niewielkie modyfikacje architektur BERT lub GPT, z analogicznym pretreningiem.

g) Główne cechy szczególne architektury PatchTST to operowanie niezależnie na jednowymiarowych szeregach (ang. *channel independence*), współdzielenie parametrów (ang. *weight sharing*) oraz tokenizacja (ang. *tokenization / patching*).

Zadanie 10

Zaznacz prawidłowe odpowiedzi dotyczące praktycznego zastosowania szeregów czasowych.

a) Mamy pomiary sensora przemysłowego, o dużej częstotliwości i wariancji. Dane są mocno zaszumione. Sensor psuje się i czasem nie zwraca wartości. Test KPSS zwrócił wartość 0. Prognozowanie długoterminowe takiego szeregu powinno wykorzystywać imputację, resampling i transformację Box-Coxa, ale nie wymaga różnicowania.

b) Startujesz w konkursie prognozowania szeregów czasowych na platformie Kaggle. Typowo najlepszym wyborem będzie LightGBM w połączeniu ze złożoną inżynierią cech oraz uczeniem zespołowym (ang. *ensemble learning*).

c) Mamy sieć sklepów i chcemy prognozować sprzedaż dla każdego z nich. Wiemy, że zachowanie kupujących ma podobne wzorce w każdym sklepie. Dodatkowo na też silny związek ze znanymi czynnikami zewnętrznymi, jak np. niedziele, święta, promocje. W takim przypadku szczególnie korzystne będzie użycie pretrenowanego modelu neuronowego PatchTST.

d) Mamy aplikację mobilną mierzącą kroki użytkownika, tętno, jakość snu itp. i chcemy je prognozować jako szeregi czasowe. Telefon ma bardzo ograniczone zasoby sprzętowe. Ze wzgledu na potrzebę personalizacji i złozony charakter dobrym wyborem będą pretrenowane modele (ang. *pretrained model foundation models*), potrafiące wykonać *zero-shot forecasting*, jak TimeFM.

Zadanie 11

Zaznacz poprawnie zdefiniowane pojęcia.

a) *Mode dropping* to problem, w którym generator generuje ograniczony zbiór bardzo podobnych próbek, ignorując pełną różnorodność danych rzeczywistych.

b) *Style Loss* to funkcja straty, badająca różnicę styli między badanym obrazem a obrazem wzorcowym. Funkcja wymaga wykorzystania pretrenowanej sieci i obliczenia macierzy Grama z odpowiednich *feature maps*.

c) *Checkerboard mask* to wzorce, które są wykorzystywane w modelach takich jak *normalizing flow* do selektywnego przetwarzania danych. Poprzez ich działanie dane (np. obrazy) są dzielone na części podlegającą dalszej modyfikacji i części pozostawioną w danym przekształceniu niezmienioną.

d) *Latent space* to przestrzeń, w której zaczynają swoje działanie generatory. Przestrzeń ta nie posiada żadnej struktury, punkty odpowiadające podobnym obiektom mogą być rozsiane losowo w przestrzeni, a punkty dla zupełnie różnych obiektów znajdować się dowolnie blisko siebie.

e) GLOW to sieć neuronowa uczona z wykorzystaniem funkcji straty *contextual loss*.

Zadanie 12

Zaznacz zdania poprawne dotyczące *reparametrization trick*.

a) Jest to metoda wykorzystywana w sieciach typu VAE.

1. Jest to metoda wykorzystywana w sieciach typu GAN.

b) Jest jednym z dostępnych wariantów przekształceń w *normalizing flow*.

c) Jest to metoda umożliwiająca zastosowanie algorytmu wstecznej propagacji błędu do próbkowania z rozkładu normalnego.

d) Jest to metoda przekształcania danych do większej liczby wymiarów.

Zadanie 13

Wskaż zdania prawdziwe w kontekście podejścia dyfuzyjnego do modeli generatywnych.

a) W czasie treningu stopniowo dodaje szum do obrazu i następnie go usuwa, aby wygenerować realistyczne obrazy.

b) Generuje obraz od razu na podstawie promptu, bez etapów przekształcania.

c) Musi korzystać z odwracalnych przekształceń pomiędzy warstwami.

d) Korzysta z sieci neuronowej do przewidywania kolejnych pikseli na podstawie sąsiednich.

e) Wykorzystuje sieć dyskryminacyjną jako kryterium treningu.

Zadanie 14

Zaznacz stacjonarne rozkłady stanów dla procesu Markowa danego następującą macierzą Markowa:

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1/2 & 1/2 \\ 0 & 3/4 & 1/4 \end{bmatrix}$$

a)

$$\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

b)

$$\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$

c)

$$\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

d)

$$\begin{bmatrix} 1/2 \\ 1/4 \\ 1/4 \end{bmatrix}$$

e)

$$\begin{bmatrix} 1 \\ 7/4 \\ -5/4 \end{bmatrix}$$

Zadanie 15

Wskaż zdania prawdziwe w kontekście modeli typu transformer.

a) BART jest modelem typu enkoder-dekoder.

b) Algorytmy z rodziny GPT wykorzystują tylko enkoder do generowania tekstu.

c) Algorytmy z rodziny GPT generują tekst w sposób autoregresyjny.

d) Transformery wykorzystują mechanizm uwagi (atencji) Bahdanau.

e) Transformery są rodzajem sieci rekurencyjnych.

Zadanie 16

Pewna choroba dotyka 2% populacji. Test na tę chorobę ma *Recall (True Positive Rate)* = 95% oraz *Specificity (True Negative Rate)* = 90%. Losowo wybrana osoba A uzyskała pozytywny wynik testu, a losowo wybrana osoba B uzyskała negatywny wynik testu. Zaznacz zdania prawdziwe.

a) Prawdopodobieństwo *a posteriori*, że osoba A jest naprawdę chora jest mniejsze od 15%.

b) Prawdopodobieństwo *a posteriori*, że osoba B jest naprawdę chora jest w przybliżeniu równe 1%.

c) Prawdopodobieństwo *a posteriori*, że osoba A jest naprawdę chora wynosi 95%.

d) Prawdopodobieństwo *a priori*, że osoba B jest zdrowa wynosi 98%.

e) Prawdopodobieństwo *a posteriori*, że osoba A jest naprawdę chora jest większe od 15%.

f) Prawdopodobieństwo *a posteriori*, że osoba B jest naprawdę zdrowa jest mniejsze od 98%.

Zadanie 17

Stosujesz Bayesowską liniową regresję wielomianową na zbiorze danych ze sporym szumem. W jaki sposób wybór stosunkowo wąskiego rozkładu Gaussa *prior* dla wag modelu (mała wariancja) wpłynie na obliczany model?

a) Zniechęci do stosowania dużych wartości wag, zapobiegając nadmiernemu dopasowaniu.

b) Zachęci do stosowania dużych wartości wag, czyniąc model bardziej elastycznym.

c) Będzie to równoważne zastosowaniu regularizacji $L2$ przy poszukiwaniu rozwiązania metodami optymalizacyjnymi.

d) Zmniejsza wariancję modelu predykcyjnego *a posteriori*.

e) Zmusza model do lepszego dopasowania do szumu w zbiorze danych.

Zadanie 18

Zaznacz zdania prawdziwe.

a) Procesy Gaussowskie wykorzystują nieskończenie wymiarową reprezentację funkcji bazowych.

b) Liniowa regresja Bayesowska zakłada stałą liczbę funkcji bazowych.

c) Regresja procesem Gaussowskim zwraca pełny rozkład predykcyjny, co oznacza, że dostajemy zarówno średnią wartość predykcji, jak i jej wariancję.

d) W bayesowskiej regresji liniowej można całkowicie pominąć założenia o rozkładzie danych, bo model sam się ich nauczy.

e) Regresja Bayesa zawsze daje dokładniejsze wyniki niż regresja liniowa GP.

Zadanie 19

Które z poniższych stwierdzeń na temat standardowej wersji algorytmu EM (ang. *Expectation-Maximization*) w kontekście uczenia niezadzorowanego i maksymalizacji funkcji wiarygodności są poprawne?

a) Algorytm EM gwarantuje znalezienie globalnego optimum funkcji wiarygodności.

b) W kroku E (ang. *e-step*) algorytm aktualizuje parametry modelu, wykorzystując w tym celu estymację funkcji wiarygodności (ang. *likelihood function*).

c) Algorytm EM iteracyjnie maksymalizuje funkcję wiarygodności poprzez naprzemienne szacowanie odpowiedzialności (ang. *responsibility*) i aktualizację parametrów modelu.

d) Algorytm EM wymaga znajomości prawdziwych etykiet punktów danych z wyprzedzeniem.

e) Algorytm EM może skutecznie usuwać składniki mikstury Gaussowskiej, przypisując im współczynniki mieszania (ang. *mixing coefficients*) bliskie zeru.

Zadanie 20

Które z poniższych metod mogą być użyte do określenia optymalnej liczby składników w modelu *Gaussian Mixture Model (GMM)*?

a) Akaike Information Criterion (AIC).

1. Walidacja krzyżowa (ang. *cross-validation*).

b) Zastosowanie modeli mieszanych z rozkładem Bernoulliego.

c) Zastosowanie algorytmu ARD (ang. *Automatic Relevance Determination*).

d) Zastosowanie modeli rozszerzonych takich jak *Bayesian Gaussian Mixture Model*.

Klucz prawidłowych odpowiedzi

Nr pytania	Prawidłowe odpowiedzi
1	b, d
2	b, d
3	a, e
4	b, c, e
5	a, c, f, g, i, k
6	b, d, e, f
7	a, b, c, d, e, f
8	g
9	a, b, d, e, g
10	a, b
11	a, b, c
12	a, d
13	a
14	a, d
15	a, c
16	b, d, e
17	a, c, d
18	a, b, c
19	c, e
20	a, b, e