COME OTTENERE LA CERTIFICAZIONE AWS MACHINE LEARNING – SPECIALTY IN 10 GIORNI DA DEVOPS ENGINEER.

APN Training Partner | AWS Training and Certification | ML

Alessandro Gaggia | 10 Luglio 2020

Come ricorderete, qualche giorno fa abbiamo festeggiato ***la sesta (!) certificazione ottenuta dal nostro Cloud Expert Alessandro Gaggia****: l'****AWS Certified Machine Learning Specialty****, la 58° per beSharp. Alessandro è una colonna portante di beSharp: è entrato a far parte del team come Front-end developer nel lontano 2012, a pochi mesi dalla nascita di beSharp, e oggi è il decano del team di sviluppo. Avvicinatosi al Machine Learning solamente un anno fa per partecipare all'AWS DeepRacer League tenutasi durante l'AWS RE:Mars a Las Vegas, ne ha fatto una vera e propria passione. Una volta tirato il fiato dopo il rush verso la certificazione (e non prima di un brindisi celebrativo virtuale col resto del team), Alessandro si è buttato in una preziosa retrospettiva sul percorso vincente che lo ha portato ad ottenere in pochissimi giorni questo invidiabile risultato. È il momento quindi di lasciargli la parola. Pronti a diventare AWS Certified?*

La mia esperienza

L'**AWS Certified Machine Learning – Speciality** è una certificazione avanzata un po' diversa dalle altre, in quanto è l'unica a richiedere una conoscenza specifica di settore che va oltre a quelle strettamente legate ai servizi di AWS. Infatti, per passare l'esame ed ottenere questa certificazione, è fondamentale **saper riconoscere, analizzare ed ottimizzare diversi problemi di Machine Learning** a partire dalla descrizione di casi d'uso, senza che questi siano necessariamente legati a particolari soluzioni AWS.

Questo mio punto di vista si discosta da quanto ho potuto leggere in molti articoli, in cui diversi autori considerano SageMaker come focus principale. Sono convinto, invece, che questo esame abbia testato soprattutto la mia dimestichezza con il Machine Learning in generale.

Dall'altro lato, numerose testimonianze – per lo più di data scientist esperti – affermano che prima di affrontare questa certificazione siano necessari diversi anni di esperienza sul campo.

Personalmente, trovo che non sia strettamente necessario: l'esperienza nel settore sicuramente gioca un ruolo importante tuttavia, come per tutte le altre certificazioni AWS che ho conseguito, serve principalmente dedizione nello studio, desiderio di approfondire gli argomenti e capacità logica per riuscire nell'impresa. Anche avere qualche nozione di Big Data può aiutare.

Disclaimer: le informazioni qui raccolte rappresentano la mia personale esperienza per la preparazione dell'esame e non vogliono porsi come un sostituto, tantomeno esaustivo, del materiale indicato da Amazon Web Services. Credo però che rappresentino un ottimo spunto da seguire per ottimizzare il proprio percorso di studio, contenendo riferimenti a diversi siti sull'argomento e alle FAQ di AWS.

Particolare attenzione va anche posta ai **diversi questionari di prova** disponibili, di cui verrà fornito il link, ed in particolar modo all'eseguibile di Certbolt, che permette di simulare un test completo di 65 domande con soluzione.

Con una full-immersion di studio da 8 ore al giorno, la certificazione può essere preparata in circa 10 giorni… o almeno, questa è stata la mia strategia, che mi ha permesso di passare l'esame con uno score di 800/1000 su un minimo richiesto di 750/1000.

Gli argomenti dell'esame

L'esame dura **170 minuti per 65 domande**, ma personalmente ho consegnato dopo solo 90 minuti. Si ha quindi il tempo per riguardare tutto con calma. Una nota positiva è che, rispetto ad altre certificazioni advanced, questa ha **casi d'uso corti e veloci da leggere**, e questo fa guadagnare tempo.

Gli argomenti principali che ho ritrovato sia nelle domande di prova che in quelle ufficiali sono a grandi linee i seguenti:

- Dato un problema di ML, utilizzare una combinazione di servizi managed di AWS per risolverlo nella maniera più veloce ed efficiente possibile. Questo caso è coperto dai seguenti topic:
  - **Kinesis Data Stream, Analytics, Firehose.**
  - **S3, KMS, DynamoDB** (in modo blando, sapere cosa sono e cosa ci puoi fare)
  - **AWS ElasticSearch con Kibana, Splunk, AWS Quicksight** (per la Business intelligence – di seguito abbreviata in BI)
  - **AWS Athena, AWS Redshift**
  - **AWS EMR con Spark**
  - **AWS Glue**
  - **AWS Translate, Transcribe, Lex, Polly, DeepLens, Rekognition, Comprehend, etc.**

In alcuni casi si ha anche una combinazione di Managed Services e Sagemaker, ma in questi casi si riesce a capire più facilmente come escludere le risposte sbagliate.

- Dato un problema di ML, gestire le seguenti casistiche:

- Pulizia dei dati con tecniche appropriate quali: **Filtering, Transforming, Scaling, Extracting**
- Gestire **Overfitting/Underfitting** e capire in quale caso stiamo agendo su elementi quali: **Learning Rate, Batch Size, Oversampling, Regularization, Dropout Rate, Feature Increasing, Denoising, Normalization, Epoch Time, etc.**
- Valutare lo score più appropriato per un modello in base alla sua tipologia e alle richieste di Business; imparare a distinguere tra: **RMSE, R2, F1, Recall, Accuracy, Precision, AUS-ROC, etc.**
- Gestire il tuning degli **Hyperparameters** sia su Sagemaker che in generale; vedere dunque: **Learning Rate, Epoch Time, L1 e L2 Regularization, Adam, Random e Stocastic Optimization, parametri alpha, beta e gamma** in diversi contesti.
- Gestire il deploy su Sagemaker

Se vogliamo analizzare gli argomenti più ad alto livello, alcuni "must" sono:

- Creare delle pipeline usando Kinesis Stream, Firehose e Analytics usandoli in combinazione con
- Amazon Athena o Elastic Search
- Risolvere un problema di Overfitting
- Risolvere un problema di Tuning
- Risolvere un problema di Scoring
- Risolvere un problema di Confusion matrix
- Usare una combinazione di servizi ML managed di AWS per risolvere un problema in modo facile e veloce
- **Gestire i diversi tipi di deploy su Sagemaker**

Questa carrellata copre a grandi linee il possibile ventaglio di domande che ci si può trovare ad affrontare all'esame. Qui di seguito descrivo una scaletta di studio, tramite la quale è possibile affrontare gli argomenti in modo più ordinato ed organico.

Gli argomenti da studiare: una scaletta

I link di riferimento agli argomenti trattati sono in gran parte relativi alla documentazione ufficiale di AWS, in particolar modo per SageMaker e i diversi servizi managed. Inoltre, per quanto riguarda i temi più strettamente legati al Machine Learning, ho voluto proporre alcuni siti che ho trovato particolarmente completi, ben spiegati ed esaustivi.

- Cos'è il machine learning: come si differenzia da Artificial Intelligence e Deep Learning: https://machinelearningmastery.com/machine-learning-for-programmers/ https://www.geeksforgeeks.org/difference-between-artificial-intelligence-vs-machine-learning-vs-deep-learning/?ref=rp

- La Pipeline di ML in generale e per AWS: https://medium.com/slalom-data-analytics/mlops-part-2-machine-learning-pipeline-automation-with-aws-1ca10348239e
- Come si procede a gestire un problema di ML: https://towardsdatascience.com/how-to-approach-a-machine-learning-problem-3fe843fd1166

https://towardsdatascience.com/task-cheatsheet-for-almost-every-machine-learning-project-d0946861c6d0

- Machine Learning su AWS intro: Servizi Managed (concentrarsi su che servizi offrono)

- AWS GLUE: https://aws.amazon.com/it/glue/faqs/
- AWS DATAPIPELINE: https://aws.amazon.com/it/datapipeline/faqs/
- AWS AUGMENTED AI: https://aws.amazon.com/it/augmented-ai/faqs/
- AWS DEEP LENS: https://aws.amazon.com/it/deeplens/faqs/
- AWS TRANSLATE: https://aws.amazon.com/it/translate/faqs/
- AMAZON TRANSCRIBE: https://aws.amazon.com/it/transcribe/faqs/
- AMAZON TEXTRACT: https://aws.amazon.com/it/textract/faqs/
- AWS REKOGNITION: https://aws.amazon.com/it/rekognition/faqs/
- AWS POLLY: https://aws.amazon.com/it/polly/faqs/
- AWS PERSONALIZE: https://aws.amazon.com/it/personalize/faqs/
- AMAZON LEX: https://aws.amazon.com/it/lex/faqs/
- AMAZON KENDRA: https://aws.amazon.com/it/kendra/faqs/
- AMAZON FORECAST: https://aws.amazon.com/it/forecast/faqs/
- AMAZON COMPREHEND: https://aws.amazon.com/it/comprehend/faqs/
- ML Supervised: https://towardsdatascience.com/types-of-machine-learning-algorithms-you-should-know-953a08248861

  - Linear Regression
  - Forecasting
  - Classification

- ML Unsupervised: https://www.guru99.com/unsupervised-machine-learning.html

  - Clustering
  - Anomaly Detection
  - Topic modeling
  - Machine translation
  - Reinforcement Learning

- Sagemaker Ground Truth: https://aws.amazon.com/it/sagemaker/groundtruth/faqs/
- Il concetto di Datalake e S3: https://aws.amazon.com/it/big-data/datalakes-and-analytics/what-is-a-data-lake/?nc=sn&loc=2
- Pulizia e formattazione dei Dati

  - I dati categorici e nominali

    - I dati sporchi
    - Gli outliers
    - I dati mancanti
    - I grafici di Pandas
    - Correlation Matrix
    - Overfitting e Underfitting
    - Splitting dei dati
    - Simple hold out
    - K Fold
    - K Fold random shuffle
    - Stratified K Fold
    - Leave one out

- Strumenti di analisi per Supervised/Unsupervised

    - ML Supervised: https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234

        - RMSE, MSE, R2
        - Accuracy, F1 Score, Precision, Recall, TNR, AUC-ROC

    - ML Unsupervised: https://www.guavus.com/technical-blog/unsupervised-machine-learning-validation-techniques/

        - Validazione interna
        - Validazione esterna
        - Twin-Sample Validation

    - ML Deep Learning: qualche nozione

- Sagemaker: Training Job come gestirlo e come creare le configurazioni: https://docs.aws.amazon.com/sagemaker/latest/dg/train-model.html
- Sagemaker: i tipi di soluzioni possibili
- Sagemaker: gli algoritmi nel dettaglio: https://docs.aws.amazon.com/sagemaker/latest/dg/algos.html

    - Blazing Text
    - Deep AR
    - Factorization Machine
    - Image Classification
    - IP Insight

    - K-Means
    - KNN
    - LDA
    - Linear Learner
    - Neural Topic Model
    - Object2Vect
    - Object Detection
    - Principal Component analysis
    - Random Cut Forest
    - Semantic Segmentation
    - Seq2Seq
    - XGBoost

- Sagemaker: il tuning: https://towardsdatascience.com/demystifying-model-training-tuning-f4e6b46e7307

    - Feature extraction
    - Numeric Transformation
    - Binning
    - Scaling
    - Dati categorici o nominali
    - Agiamo sui parametri
    - Agiamo sugli hyperparameter

- Sagemaker: il deploy: https://docs.aws.amazon.com/sagemaker/latest/dg/deploy-model.html

    - Deploy Autogestito
    - Deploy mediante Sagemaker
    - Come deployare

        - Deploy Blue/Green: https://docs.aws.amazon.com/whitepapers/latest/wellarchitected-machine-learning-lens/bluegreen-deployments.html
        - Deploy A/B:
        - https://docs.aws.amazon.com/whitepapers/latest/wellarchitected-machine-learning-lens/ab-testing.html
        - Canary: https://docs.aws.amazon.com/whitepapers/latest/wellarchitected-machine-learning-lens/canary-deployment.html

- Batch Inference: https://docs.aws.amazon.com/sagemaker/latest/dg/batch-transform.html
- Online Inference: https://docs.aws.amazon.com/sagemaker/latest/dg/inference-pipeline-real-time.html
- Online Vs Batch: https://mlinproduction.com/batch-inference-vs-online-inference/
- Sagemaker: il logging e il Concept Drift: https://docs.aws.amazon.com/sagemaker/latest/dg/monitoring-overview.html
- Kinesis Data Stream VS Firehose: https://aws.amazon.com/it/kinesis/data-streams/faqs/?nc=sn&loc=5
  https://aws.amazon.com/it/kinesis/data-firehose/faqs/

Cheats, tips & tricks

1. Generalmente, nel caso di **Simple Hold Out,** i valori tipici sono: **80/10/10** oppure **70/15/15**.
2. Se nelle domande vengono introdotti i termini **historical data** allora si avrà a che fare con un algoritmo **supervised**.
3. Che vantaggio portano i **Random** e Bayesian **Optimizer** per gli hyperparameters in Sagemaker rispetto al **Grid Optimizer o al manual? Esplorano meglio lo spazio dei parametri verificando meglio combinazioni inusuali.**
4. Random optimizer è più veloce di Bayesian, ma il secondo è più preciso
5. Una confusion matrix può anche essere **NxN**.
6. Term Frequency – Inverse Document Frequency: un valore alto indica un termine raro
7. Oversampling e undersampling **non hanno utilità per le regression (sia logistic che linear)**
8. Quando ci troviamo di fronte ad un problema con una classe fortemente sbilanciata, la **class probability threshold può essere portata da 0.5 a un valore maggiore a favore della classe sbilanciata.**
9. Oversampling **va effettuato dopo lo splitting dei dati per evitare fenomeni di bleeding dei dati.**
10. Per **uscire dai minimi locali** devo **ridurre il batch size** e **ridurre anche il learning rate per smorzare l'effetto oscillatorio di un batch size piccolo e favorire la convergenza.**
11. Per fare oversampling la logica è **GAN migliore di SMOTE migliore di manuale.**
12. Se abbiamo problemi di missing value e **una possibile soluzione proposta è il machine learning per rimpiazzare i valori**, in genere è la scelta migliore.
13. Quando ho problemi di **recommendation engine o in genere di predire scelte di utenti** se abbiamo tra le scelte disponibili **collaborative filtering**, è quasi sempre una ottima scelta.
14. Se dobbiamo scegliere: **AWS Glue è batch oriented, Kinesis Firehose è per streaming real time data.**
15. Se abbiamo problemi di forecasting di clima, usiamo generalmente la **Regressione Lineare (se non presenti algoritmi più fitting, ma in genere è sufficiente)**.
16. La Lambda di input di Firehose ha un timeout di **3 sec, quindi se eseguiamo operazioni di input complesse di solito non è sufficiente e va aumentato.**
17. Scatter Plot si usa per analisi a 2 dimensioni, Istogramma per 1 dimensione.
18. Logistic Regression: **si usa per la classificazione binaria, non bisogna farsi ingannare dal nome!**

19. NTM: neural topic model si usa per raggruppare topic in categorie
20. Semantic Segmentation: trova feature di alto livello a partire dall'analisi dei pixel.
21. Firehose non ha senso se i dati sono già su S3, prende dati da uno stream!
22. Per il deep learning e i problemi di overfitting fare riferimento ai seguenti schemi indicativi:

    < DROPOUT + > REGULARIZATION + > FEATURE INCREASE = < OVERFITTING

    Sono i parametri da modificare per ridurre l'overfitting. Inoltre:

    > DROPOUT = > NOISINESS

23. Quando uso Kinesis Data Stream ho una variante apposita che si chiama **video stream** che si collega **in modo semplice con Rekognition Video**. E' presente un caso d'uso nelle guide di AWS.
24. AWS Comprehend **non** fa **solo Sentiment Analysis** ma anche **Topic Analysis**, **Key Extraction** e **Entity Recognition**, inoltre utilizza NLP, quindi nel caso si abbiano gli stessi topic in una domanda si può riflettere su questo per capire se usare NLP.
25. L1 e L2 sono termini di regolarizzazione per ridurre l'overfitting rendendo l'algoritmo più conservativo quando si ha dataset con molte features.
26. La Cross Validation si basa sugli iperparametri, aiuta contro l'overfitting quando il dataset è piccolo, quindi quando il rischio è maggiore per la bassa quantità di dati.
27. Deep Learning: sostituisce un modello con una rete neurale per fare un'analisi
28. Nel Deep Learning le feature possono anche essere usate raw, perchè nel caso di reti complesse, sono i livelli stessi a rielaborare le feature di basso livello in nuove feature di più alto livello.
29. Nel Deep learning abbiamo diversi tipi di **activation function**:

    1. Sigmoide
    2. ReLU
    3. Softplus
    4. TanH

30. Quando si parla di Deep Learning al momento le **CNN** o Convolutional Neural Network (molto indicate per l'image processing) e le **Long Short Term Memory,** sono tra le più utilizzate per versatilità e scalabilità.

Le domande: dove trovarle?

In generale, a causa del NDA richiesto da AWS sulle domande delle certificazioni, è difficile ottenere materiale **ufficiale** (o anche solo ufficioso e affidabile)... questa lista di link punta a risorse che ho personalmente provato e verificato come funzionali alla preparazione della certificazione:

- **Domande ufficiali di AWS (gratuito):** https://d1.awsstatic.com/training-and-certification/docs-ml/AWS-Certified-Machine-Learning-Specialty_Sample-Questions.pdf

- **Esame di prova ufficiale di AWS (a pagamento):** https://www.aws.training/certification?rightcta=mlexam

- **Exam readiness class (a pagamento):** https://www.aws.training/SessionSearch?pageNumber=1&courseId=38153

- **Esame di prova di Certbolt di 65 domande** – solo per sistemi Windows – **altamente consigliato (gratuito)**: https://www.certbolt.com/aws-certified-machine-learning-specialty-exam-dumps

- **Diverso materiale proposto da Udemy (a pagamento)**: https://www.udemy.com/topic/aws-certified-machine-learning-specialty/

- Domande varie ed eventuali: cercate tramite Google scrivendo "**aws machine learning specialty dumps**". Possono essere utili per verificare la vostra conoscenza ma **sconsigliate in quanto molte di esse sono purtroppo inesatte**. Utilizzatele come controprova della vostra preparazione.

Una nota: l'esame da remoto con Pearson/VUE

L'esame da remoto prevede di accedere al pannello Pearson/VUE e registrarsi per una data in cui sia disponibile il proctoring online, necessario per accertare la validità dell'esame. Data la situazione d'emergenza di quest'anno, conviene prenotare la sessione con **largo anticipo**, per essere sicuri di trovare una data e un orario convenienti.

È necessario avere sottomano Carta d'identità o **patente** o **passaporto validi**, di cui verranno richieste delle foto chiare e comprensibili. Un documento rovinato o una foto illeggibile provocherà forti ritardi sullo scheduling previsto a causa degli accertamenti telefonici.

Verrà fatto scaricare un software disponibile per tutte le piattaforme Windows, Mac e Linux, che servirà per sostenere l'esame e che verificherà, mediante foto dal cellulare, la stanza in cui si svolgerà il test. Anche in questo caso è necessario seguire scrupolosamente le istruzioni ricevute via email alla registrazione, per effettuare i test di prova prima della data dell'esame.

Durante lo svolgimento dell'esame, un proctor vi seguirà dalla vostra webcam e il software utilizzato sarà eseguito in fullscreen per prevenire l'uso di altri programmi.

Nel caso si verificasse qualche problema, il proctor ci avviserà mediante Chat interna.

Per concludere, se vi stavate chiedendo se fosse possibile preparare l'AWS Certified Machine Learning Specialty – o una specialty AWS in generale – in autonomia, ora sapete che la risposta è... ***tecnicamente sì*** 🙂

Tuttavia, per chi fosse alle prime armi con le certificazioni AWS o volesse ricevere una formazione strutturata, ***AWS ha sviluppato i propri corsi ufficiali****, erogati tramite APN Training Partner riconosciuti.* ***beSharp è stata selezionata da Amazon Web Services nel ristretto numero di partner autorizzati a erogare l'intero catalogo di corsi ufficiali AWS****: i nostri corsi di formazione, sviluppati e gestiti da Cloud Expert pluri-certificati, garantiscono che i contenuti riflettano le best-practice più*

recenti e danno la possibilità di ricevere feedback dal vivo e risposte alle domande direttamente da un istruttore esperto. Inoltre includono ***laboratori pratici*** *pensati per consolidare le conoscenze teoriche con* ***casi d'uso reali****. Per preparare questa o qualsiasi altra certificazione ufficiale, potete leggere qui tutti i dettagli e contattarci per pianificare il vostro percorso di formazione.*

A presto, con un nuovo articolo sul nostro blog e in bocca al lupo per la vostra prossima certificazione AWS!

Alessandro Gaggia

Head of Software Development

Get in touch

beSharp.it
proud2becloud@besharp.it