Catatan Laporan

Prediksi Debit Aliran menggunakan *Long Short-Term Memory* (LSTM)

Versi 1.0.0

Berdasarkan *Jupyter Notebook*: `github_taruma_demo_lstm_rr_catatan.ipynb`

oleh Taruma Sakti Megariansyah

22 Oktober 2019

github.com/taruma/vivaldi

Dokumen ini merupakan catatan untuk laporan "Prediksi Debit Aliran Menggunakan Metode *Long Short-Term Memory* (LSTM)" atau berkas `github-taruma_demo_lstm_rr.ipynb`.

1 Info Dataset

Dataset beserta informasinya diperoleh dari skripsi saya sendiri berjudul "Kajian Penerapan Model NRECA di Bendung Pamarayan" pada tahun 2015. Data curah hujan dan debit diperoleh dari skripsi. Untuk data klimatologi, diunduh melalui Data Online BMKG yang diakses pada 2 Oktober 2019, dikarenakan data dari stasiun terdekat tidak lengkap. Saya akan mengusahakan menyertakan segala informasi mengenai dataset yang perlu diketahui di dalam catatan ini.

1.1 Dataset

Dataset merupakan data hidrologi dan klimatologi **harian** dari tanggal **1 Maret 1998** sampai **31 Desember 2008** (3959 hari). Dataset terpisah menjadi 3 kategori yaitu: data curah hujan, data klimatologi, dan data debit.

- Data curah hujan diperoleh dari 8 stasiun yaitu: `bojong_manik`, `gunung_tunggal`, `pasir_ona`, `sampang_peundeuy`, `cimarga`, `bd_pamarayan`, `ciminyak_cilaki`, `gardu_tanjak`.
- Data debit diperoleh dari 1 stasiun yaitu: `bd_pamarayan`.
- Data klimatologi diperoleh dari 1 stasiun yaitu: `geofisika_serang`.

1.2 Sumber Dataset

Berikut sumber dataset yang diperoleh (Megariansyah, 2015):

- Data Curah Hujan, 8 Stasiun: BBWS Cidanau-Ciujung-Cidurian
- Data Debit, 1 Stasiun: BBWS Cidanau-Ciujung-Cidurian
- Data Klimatologi, 1 Stasiun: Data Online BMKG

1.3 Ringkasan Dataset

- Data curah hujan merupakan data berkolom tunggal yang menunjukkan besarnya curah hujan dalam satuan *mm* untuk masing-masing stasiun.
- Data debit merupakan data berkolom tunggal yang menunjukkan besarnya debit dalam satuan m^3/s.
- Data klimatologi merupakan data dengan 10 kolom berupa:
  - Arah angin saat kecepatan maksimum (ddd_x) dalam satuan derajat
  - Arah angin terbanyak (ddd_car) dalam satuan derajat
  - Curah hujan (RR) dalam satuan *mm*
  - Kecepatan angin maksimum (ff_x) dalam satuan m/s
  - Kecepatan angin rata-rata (ff_avg) dalam satuan m/s
  - Kelembapan rata-rata (RH_avg) dalam satuan %
  - Lamanya penyinaran matahari (ss) dalam satuan jam
  - Temperatur maksimum (Tx) dalam derajat Celcius
  - Temperatur minimum (Tn) dalam derajat Celcius
  - Temperatur rata-rata (Tavg) dalam derajat Celcius
- Data debit merupakan variabel dependen, sedangkan data lainnya merupakan variabel independen.

- Pada data klimatologi, isian yang bernilai `8888` berarti data tidak diukur dan isian yang bernilai `9999` berarti tidak ada data (tidak dilakukan pengukuran). Nilai tersebut akan dianggap nilai yang hilang "`NaN`".

2 Strategi Penyelesaian

Terdapat 5 tahap yang saya ikuti dalam menjawab objektif buku ini.

1. Tahap 0: Pengaturan Awal dan Inisiasi

   Pada tahap ini dilakukan pengaturan awal dan inisiasi untuk mempersiapkan buku. Buku dapat dijalankan secara lokal ataupun *cloud* menggunakan Google Colab. Di tahap ini, dapat dilakukan pengaturan manual seperti penamaan buku (jika ingin dilakukan penyimpanan), menentukan lokasi dataset dan dropbox, dll.

2. Tahap 1: Akusisi Dataset

   Dataset yang diterima bisa dalam berbagai bentuk seperti dalam bentuk Excel, PDF, bahkan fisik berupa lembaran/laporan. Pada tahap ini dilakukan pengubahan dataset tersebut biar bisa diolah secara digital. Untungnya, pada buku ini, dataset yang diperoleh berupa digital dengan format .xls sehingga memudahkan dalam mempersiapkan pengolahan data lebih lanjut.

   Untuk membantu tahap ini juga dibuat modul khusus yang telah tersedia di hidrokit yang dapat diakses melalui `hidrokit.contrib.taruma` dengan nama modul `hk43` untuk data hujan/debit dan `hk73` untuk data klimatologi/bmkg.

3. Tahap 2: Prapemrosesan Data

   Tahap ini memastikan kelengkapan data dan validitas data. Prapemrosesan dapat berupa mencari nilai invalid dan mengoreksinya, memeriksa data yang hilang dan dikoreksi dengan berbagai metode (pada buku ini menggunakan interpolasi linear). Karena pemodelan bergantung dengan data yang digunakan, tahap ini memiliki peran penting dalam keberhasilan pemodelan.

4. Tahap 3: Input Pemodelan

   Data yang telah melewati tahap prapemrosesan akan dipersiapkan untuk digunakan dalam pemodelan. Persiapan ini berupa memisahkan dataset menjadi dua bagian yaitu *train set* dan *test set*, normalisasi, dan transformasi dataset.

   Pada pemodelan *Recurrent Neural Networks*, input yang diterima berbentuk tensor 3D. Pada manual Keras, disebutkan bahwa dimensi tensor 3D berupa (*batch_size*, *timesteps*, *input_dim*).

   Dalam buku ini, digunakan `TIMESTEPS=365` hari serupa pada makalah Kratzert et. al. (2018). Nilai *timesteps* tidak harus bernilai `365` di buku ini, nilai *timesteps* dapat di isi dengan nilai sembarang sampai memperoleh nilai optimal untuk model.

   Untuk membantu tahap ini, dibuat modul khusus yang dapat diakses melalui `hidrokit.contrib.taruma.hk53`.

5. Tahap 4: Melatih Model

Pada tahap ini harus ditentukan arsitektur RNN/LSTM yang akan digunakan. Parameter seperti jumlah *hidden layer*, jumlah *units*, penggunaan *dropout layer*, jenis aktivasi, dll.

Untuk menyederhanakan permasalahan, penggunaan parameter selain yang disebutkan dibawah ini menggunakan nilai `default` dari program:

- optimizer: `adam`
- activation: `sigmoid`
- probability dropout: `0.1`
- units: `20`/lstm-layer
- loss function: `mean squared error`
- epoch: `50`
- batch_size: `30`

Di tahap ini juga dibuat fungsi khusus untuk memperoleh metrik setiap epoch. Fungsi khusus yang dibuat antara lain `nse`, `nse_mod`, dan `r_squared`.

6. Tahap 5: Evaluasi Model

   Evaluasi yang dilakukan antara lain: melihat perkembangan metrik pada setiap epoch, mengevaluasi *train set* dan *test set*.

   Penilaian performa model bergantung pada metrik yang dihasilkan oleh *test set*. Metrik yang digunakan sebagai penilaian yaitu *mean squared error* (*loss function*), *mean absolute error*, *Nash-Sutcliffe Efficiency*, *Modified NSE*, dan *Coefficient of Determination*.

3 Daftar Pustaka

Berikut daftar pustaka yang berkaitan dengan laporan "Prediksi Debit Aliran menggunakan *Long Short-Term Memory* (LSTM)". Untuk daftar pustaka yang saya gunakan bisa dilihat pada halaman github.com/taruma/vivaldi.

3.1 Dataset

- Megariansyah, Taruma S. (2015): Kajian Penerapan Model NRECA di Bendung Pamarayan, Skripsi Program Sarjana, Universitas Katolik Parahyangan.
- BMKG (2019): Data Online BMKG, diperoleh melalui situs internet: dataonline.bmkg.go.id (diakses pada: 2 Oktober 2019).

3.2 Makalah / Laporan

- Kratzert, F., Klotz, D., Brenner, C., Schulz, K., Herrnegger, M., 2018. Rainfall–runoff modelling using Long Short-Term Memory (LSTM) networks. Hydrology and Earth System Sciences 22, 6005–6022. https://doi.org/10.5194/hess-22-6005-2018
- LeCun, Y. A., Bottou, L., Orr, G. B., and Müller, K. R.: Efficient backprop, Springer, Berlin, Heidelberg, Germany, 2012.
- Minns, A. W. and Hall, M. J.: Artificial neural networks as rainfall- runoff models, Hydrolog. Sci. J., 41, 399–417, 1996.

3.3 Program

3.3.1 Bahasa Pemrograman

- Van Rossum, G. & Drake Jr, F.L., 1995. *Python tutorial*, Centrum voor Wiskunde en Informatica Amsterdam, The Netherlands.

3.3.2 Paket Scipy (Scientific Computing in Python)

- Fernando Pérez and Brian E. Granger. IPython: A System for Interactive Scientific Computing, Computing in Science & Engineering, 9, 21-29 (2007), DOI:10.1109/MCSE.2007.53
- John D. Hunter. Matplotlib: A 2D Graphics Environment, Computing in Science & Engineering, 9, 90-95 (2007), DOI:10.1109/MCSE.2007.55
- Stéfan van der Walt, S. Chris Colbert and Gaël Varoquaux. The NumPy Array: A Structure for Efficient Numerical Computation, Computing in Science & Engineering, 13, 22-30 (2011), DOI:10.1109/MCSE.2011.37
- Wes McKinney. Data Structures for Statistical Computing in Python, Proceedings of the 9th Python in Science Conference, 51-56 (2010)

3.3.3 Jupyter Notebook

- Thomas, K., Benjamin, R.-K., Fernando, P., Brian, G., Matthias, B., Jonathan, F., ... Team, J. D. (2016). Jupyter Notebooks – a publishing format for reproducible computational workflows. Stand Alone, 87–90. https://doi.org/10.3233/978-1-61499-649-1-87

3.3.4 Paket Deep Learning

- Abadi, Mart'in et al., 2016. Tensorflow: A system for large-scale machine learning. In 12th *USENIX* Symposium on Operating Systems Design and Implementation (*OSDI* 16). pp. 265–283.
- Chollet, F.: Keras, available at: https://github.com/keras-team/keras, 2015.

3.3.5 Paket Python

- Megariansyah, Taruma. (2019, October 15). hidrokit: Analisis Hidrologi dengan Python (Version 0.3.2). Zenodo. http://doi.org/10.5281/zenodo.3490672
- Roberts, W., Williams, G., Jackson, E., Nelson, E., Ames, D., 2018. Hydrostats: A Python Package for Characterizing Errors between Observed and Predicted Time Series. Hydrology 5(4) 66, doi:10.3390/hydrology5040066

4 Daftar Pranala

Berikut daftar pranala yang disinggung pada laporan.

4.1 Pranala Buku

- Google Colab: https://colab.research.google.com/drive/1bx3ak_20dcJ7VdGR-djysLIxLaX7pRI2
- Github: https://github.com/taruma/vivaldi/blob/master/notebook/github_taruma_demo_lstm_rr.ipyn

- NBViewer: https://nbviewer.jupyter.org/github/taruma/vivaldi/blob/master/notebook/github_taruma
- Laporan: https://github.com/taruma/vivaldi/blob/master/pdf/taruma_lstm_rr_laporan.pdf
- Laporan (Rapih): https://github.com/taruma/vivaldi/blob/master/pdf/taruma_lstm_rr_laporan_rapih.
- Catatan: https://github.com/taruma/vivaldi/blob/master/pdf/taruma_lstm_rr_catatan.pdf

4.2 Catatan

- Google Colab: https://colab.research.google.com/
- taruma/vivaldi: https://github.com/taruma/vivaldi

4.3 Panduan

- hk43: https://nbviewer.jupyter.org/gist/taruma/a9dd4ea61db2526853b99600909e9c50
- hk73: https://nbviewer.jupyter.org/gist/taruma/b00880905f297013f046dad95dc2e284
- hk53: https://nbviewer.jupyter.org/gist/taruma/50460ebfaab5a30c41e7f1a1ac0853e2

4.4 Tulisan

- https://towardsdatascience.com/smarter-ways-to-encode-categorical-data-for-machine-learning-part-1-of-3-6dca2f71b159
- https://towardsdatascience.com/basic-feature-engineering-to-reach-more-efficient-machine-learning-6294022e17a5

4.5 Referensi

- https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html
- https://keras.io/layers/recurrent/
- https://en.wikipedia.org/wiki/Nash–Sutcliffe_model_efficiency_coefficient

4.6 HydroStats / HydroErr

- https://github.com/BYU-Hydroinformatics/HydroErr
- https://github.com/BYU-Hydroinformatics/Hydrostats

4.7 LICENSE

- MIT: https://github.com/taruma/vivaldi/blob/master/LICENSE
- CC-BY-4.0: https://creativecommons.org/licenses/by/4.0/

5 Referensi Belajar

Saya tidak memiliki latar belakang dalam bidang komputer sehingga apa yang saya pelajari murni dari belajar otodidak yang materi pembelajarannya diperoleh daring *online*. Jika tertarik daftar materi pembelajaran saya, bisa lihat di profil koding saya di taruma.github.io/koding (akan saya perbarui dengan daftar yang lengkap jika sempat. hehe). Saya hanya akan menyebutkan beberapa kelas/kursus yang bermanfaat dalam pembuatan buku ini.

5.1 Belajar Python

Saya mempelajari python dimulai dari akhir tahun 2017 sampai sekarang, jadi masih tergolong awam juga. Saya memulai belajar python dengan ketertarikan dalam dunia *data science*. Saya mengambil kelas yang tersedia gratis (*audit access*) di edX.org. Berikut beberapa kelas yang saya ambil:

- [edX] Introduction to Python (DEV236x, DEV274x, DEV330x) oleh Microsoft.
- [edX] Data Science Research Method: Python Edition (DAT273x) oleh Microsoft.
- [udemy] Python for Data Science and Machine Learning Bootcamp oleh Jose Portilla.
- [edX] Using Python for Research oleh HarvardX.

5.2 Belajar Machine Learning / Deep Learning

Berikut beberapa kelas yang saya ambil terkait deep learning:

- [edX] Data Science Essentials (DEV203.1x) dan Principle of Machine Learning (DEV203.2x) oleh Microsoft.
- [udemy] Machine Learning A-Z™: Hands-On Python & R In Data Science oleh Kirill Eremenko, Hadelin de Ponteves, SuperDataScience.
- [udemy] Deep Learning A-Z™: Hands-On Artificial Neural Networks oleh Kirill Eremenko, Hadelin de Ponteves, SuperDataScience.

5.3 Video Youtube

Selain dari kelas juga saya menonton materi dari video youtube. Berikut daftar video youtube yang membantu saya mempelajari python/machine learning (Judul video/playlist oleh @nama channel):

- Python Tutorial oleh @Corey Schafer.
- Python Tutorial (Machine Learning with Python, Deep Learning basics with Python, Tensorflow and Keras) oleh @sentdex.
- Data Science with Python Pandas by Athena Kan oleh @CS50.
- Roadmap: How to Learn Machine Learning in 6 Months by Zach Miller oleh @IDEAS.

6 Changelog

```
- 20191022 - 1.0.0 - Initial
```