تمرین سری ۳ واحد درسی داده کاوی

جناب آقای دکترفراهانی
دستیاران آموزشی : نوید کاشی ، علی شریفی

۳ خرداد ۱۴۰۰

توجه کنید شما میتوانید بر روی کگل یا کولب و یا کامپیوتر های شخصی خود کار کنید .
به جای دانلود و آپلود دیتاست در گوگل درایو برای استفاده در کولب میتوانید به شیوه زیر عمل کنید .
چگونه از دیتاست های کگل در کولب استفاده کنیم ؟
ددلاین تمرین تا تاریخ ۲۰ روز بعد از امتحان درس داده کاوی مقطع کارشناسی ارشد میباشد .

۱ تمرین ها

۱. در خصوص کرنل های پرکاربرد روش SVM تحقیق کنید .به صورت کلی چرا ما از ایده کرنل در بحث SVM بهره میبریم . آیا میتوان در خصوص کرنل ها و استفاده ی آنها حکم کلی داد . به طور مثال بگوییم از کرنل RBF در این مواقع خاص استفاده میکنیم .

۲. قبلا با دیتاست کلاس بندی قیمت موبایل در کگل کار کرده ایم . بر روی دیتاست ، روش SVM را اجرا کنید . (استفاده از پکیج ها همانند sklearn مجاز است .)

۳. برای سوال ۲ حداقل ۵ حالت مختلف از قبیل کرنل ها و پارامترها را بررسی کنید و نتایج آن را گزارش دهید .

۴. برای سوال ۲ سعی کنید مبحث soft margin و hard margin را بررسی کنید و نتایج آن را گزارش دهید .

۵. مهندسی ویژگی یکی از بخش های مهم در فرایندهای علم داده میباشد . بر روی دیتاست موارد زیر را اجرا کنید .

(آ) بر روی فیچر battery power از روش binning استفاده کنید . (حداقل سه اندازه مختلف برای بین ها در نظر بگیرید و حتی سایز بین ها را نامساوی در نظر بگیرید .)

(ب) بر فیچرهای کتگوریکال در دیتاست one hot encoding را اعمال کنید . چرا ما باید به صورت کلی از این کدگذاری بهره ببریم .

(ج) بررسی کنید آیا استفاده از تبدیل هایی از قبیل log transform و یا تبدیل نمایی در اینجا کاربرد دارد . به صورت کلی چرا از این دست تبدیلات بهره میبریم . (در این بخش شما مجاز هستید اگر تبدیل دیگری را مناسب میدانید اعمال کنید این بخش نمره امتیازی برای شما خواهد داشت . حتما دلیل استفاده از تبدیل استفاده شده را بیان کنید .)

(د) یک فیچر جدید به نام مساحت یا حجم گوشی بسازید .

۶. برای هریک از حالت های سوال ۵ یک مدل SVM بسازید و بررسی کنید یکبار هم هر ۵ حالت را باهم اعمال کنید و مدل SVM روی آنها اجرا کنید . حاصل این مدل ها را گزارش کنید .

۷. در خصوص الگوریتم های مختلف ساخت درخت تصمیم (همانند ID3 ، CART و...) تحقیق کنید . به صورت کلی تفاوت الگوریتم های مختلف ساخت درخت تصمیم در چیست ؟

۸. به دلخواه با استفاده از پکیج ها بر روی دیتاست مطرح شده یک درخت تصمیم بسازید .

۹. برای درخت تصمیم پارامتر های مختلف مورد ارزیابی قرار دهید . آیا عمق درخت و تعداد نمونه های موجود در هر هر گره تاثیری در عملکرد درخت تصمیم دارد ؟

۱۰. در خصوص هرس کردن Pruning درخت تصمیم تحقیق کنید . چرا ما به بحث هرس کردن درخت تصمیم نیاز دارد و چه کمکی به ما میکند .

۱۱. (**بخش امتیازی**) سعی کنید این هرس کردن درخت در مدل خود اجرا کنید و بررسی کنید آیا این هرس کردن در نتایج شما تاثیر داشته است .

۱۲. بر روی دیتاست یک random forest اجرا کنید و نتایج با یک درخت تصمیم مقایسه کنید . بررسی کنید آیا نتایج درخت تصمیم بهتر است و یا random forest و چرا ؟

۱۳. تحقیق کنید چرا با وجود روش های جدید از قبیل یادگیری عمیق و شبکه عصبی ، هم چنان روشی مانند درخت تصمیم محبوب است ؟ (راهنمایی : میتوانید در خصوص بحث تفسیری پذیری این مدل ها صحبت کنید .)

۱۴. (**بخش امتیازی**) در درخت های تصمیم ما قوانینی استخراج میکنیم و از این قوانین استفاده میکنیم. در خصوص روش های دیگری که به استخراج قوانین از روی دیتاست میپردازند (rule induction) همانند روش Ripper ، IREP ، ... تحقیق کنید و آن ها را توضیح دهید . (حداقل دو روش)

۱۵. بررسی کنید آیا از درخت های تصمیم میتوان برای حل مسائل سری زمانی استفاده کرد ؟

۱۶. به آدرس مقابل مراجعه کنید . داده های قیمت بیت کوین را از تاریخ $01-01-2010$ الی $01-05-2021$ را به صورت روزانه دریافت کنید . این کار به سادگی با تنظیم تاریخ امکان پذیر است . پس از دانلود دیتاست لازم است تغییراتی روی آن اعمال گردد که عبارتند از :

- ستون Date باید به صورتی درآید که مناسب کار کردن باشد برای این کار میتوانید از لینک مقابل استفاده کنید .
- داده های تاریخ $01-01-2010$ الی $01-01-2020$ را به عنوان داده های آموزشی در نظر بگیرید و سایر داده ها یعنی از تاریخ $02-01-2020$ الی $01-05-2021$ به عنوان داده های تست در نظر بگیرید .

۱۷. با توجه به داده های سوال قبل هدف پیش بینی قیمت بیت کوین در بازه زمانی تست میباشد . در ابتدا با مدل های ساده ای آغاز کنید . (حداقل ۱۰ مدل مختلف را بررسی کنید . مدل های کلاسیک و مدل های نوین تر مثل شبکه های عصبی همانند LSTM)

۱۸. در تمرین قبل ۳ ، ۵ تا از بهترین مدل های خود را انتخاب کنید و از ایده ensemble learning استفاده کنید . در آن از تکنیک های voting ، boosting ، bagging استفاده کنید .

۱۹. بر روی دیتاست بیت کوین ، الگوریتم AdaBoost را با پارامتر های مختلف اجرا کنید و نتایج را گزارش دهید .

۲۰. بر روی دیتاست بیت کوین ، الگوریتم RandomForest را با پارامتر های مختلف اجرا کنید و نتایج را گزارش دهید .

۲۱. لازم است برای سوالات ۱۷ الی ۲۰ ، گزارش کامل نوشته شود و با دقت این نتایج گزارش شود . نتایج را یکبار با RMSE گزارش کنید و یکبار هم فرض کنید اگر ۵ درصد خطا کنید باز هم قیمت قابل قبول است و به آن لیبل درست دهید و در غیر این صورت به آن لیبل نادرست دهید . حال دقت را گزارش دهید . یعنی به طور مثال اگر قیمت واقعی ۱۰۰ بود ما اگر در بازه ۹۵ تا ۱۰۵ هم قیمت را پیش بینی کنیم برای ما قابل قبول است و لیبل درست میگیرد در غیراین صورت به آن لیبل نادرست میدهیم .

۲۲. بر روی دیتاست بیت کوین به صورت زیر عمل کنید با استفاده مدلی مانند LSTM بررسی کنید که قیمت در گام زمانی بعد افزایشی است یا نه ؟ عملا مسئله شما به صورت یک مسئله کلاس بندی تبدیل میگردد . با چه میزان دقتی میتوانید این این روند را پیش بینی کنید .

۲۳. (بخش امتیازی) با اضافه کردن داده بخش قبل به عنوان یک فیچر ورودی سعی کنید قیمت را تخمین بزنید . آیا نتیجه کار بهتر شد ؟

۲۴. (بخش امتیازی) با بهره گیری از مقالات مختلف سعی کنید نتیجه کار را ارتقا دهید . (یکی از الگوریتم های خوب که میتواند مورد بررسی قرار گیرد XGBoost میباشد . در این بخش اگر از الگوریتم جدیدی استفاده میکنید ابتدا آن را توضیح دهید و سپس آن را گزارش کنید .) [۱]

۲۵. (بخش امتیازی) سعی کنید شاخص هایی را از روی قیمت به دست آورید . در بحث داده های بورس و صنایع وابسته شاخص هایی استخراج میشود که هرکدام بیان گر یک مطلب خاص میباشد به آنها indicator میگویند . ۵ شاخص را به دلخواه انتخاب کنید و حال سعی کنید با استفاده از قیمت در بازه های قبلی و این شاخص ها ، قیمت را تخمین بزنید . آیا این کار باعث شد که نتایج کار شما بهتر شود ؟ اگر نتایج بهتر شد چرا این اتفاق افتاده است .

۲۶. دیتاستی پیوست این فایل گردیده است که مربوط به یک شرکت تجاری موفق میباشد که مسابقه ای برگزار کرده است . این شرکت بنا بر تجربیات خود شاخص هایی را از داده ها بورسی استخراج کرده است که جهت حفظ محرمانگی ما نمی دانیم این فیچرها چگونه محاسبه شده است . به طور مثال ممکن است این فیچر یک شاخص همانند میانگین متحرک باشد . هدف ما در ابتدا استفاده از مدل های مختلف جهت پیش بینی میباشد . ابتدا بر روی دیتاست ترین مدل های خود را آموزش دهید سپس جواب های خود را با توجه به دیتاست تست تخمین بزنید . (حداقل سه مدل مختلف را بررسی کنید .)

۲۷. (بخش امتیازی) در سوال قبلی از ایده ensemble learning حداقل ۵ مدل مختلف با استفاده از تکنیک های voting و boosting و bagging بررسی کنید . (سعی کنید تا حد امکان از مدل های مختلفی برای این بخش استفاده کنید) .
بررسی کنید بین خروجی مدل های مختلف و مقادیر واقعی spearman correlation چقدر است . هم چنین بررسی کنید اگر مدل ها از نظر ساختاری شبیه هم باشند بهتر خروجی میگیریم یا مدل ها ساختار های مختلفی داشته باشند و خیلی نسبت به هم متفاوت باشند.

[۱] https://towardsdatascience.com/go-highly-accurate-or-go-home-61828afb0b13

۲ راهنمایی

برای سوال ۲۶ و ۲۷ قطعه کد زیر میتواند یک مثال باشد .

```
train data
 -------------------
                   era data_type  feature_intelligence1  ...  feature_wisdom45  feature_wisdom46  target
id                                                       ...
n000315175b67977  era1     train                   0.00  ...              0.50              0.75    0.50
n0014af834a96cdd  era1     train                   0.00  ...              0.25              1.00    0.25
n001c93979ac41d4  era1     train                   0.25  ...              0.25              0.75    0.25
n0034e4143f22a13  era1     train                   1.00  ...              1.00              1.00    0.25
n00679d1a636062f  era1     train                   0.25  ...              0.25              0.75    0.75

[5 rows x 313 columns]
test data
 -------------------
                    era   data_type  feature_intelligence1  ...  feature_wisdom45  feature_wisdom46  target
id                                                           ...
n0003aa52cab36c2  era121  validation                   0.25  ...              0.00              0.00    0.25
n000920ed083903f  era121  validation                   0.75  ...              0.50              0.50    0.50
n0038e640522c4a6  era121  validation                   1.00  ...              0.50              0.00    1.00
n004ac94a87dc54b  era121  validation                   0.75  ...              0.25              0.25    0.50
n0052fe97ea0c05f  era121  validation                   0.25  ...              0.25              1.00    0.75

[5 rows x 313 columns]
```

شکل ۱: نمایی از دیتاست ترین و تست تمرین های ۲۶ و ۲۷

```python
import pandas as pd
from xgboost import XGBRegressor

# training data contains features and targets
training_data = pd.read_csv("training_data.csv").set_index("id")

# tournament data contains features only
tournament_data = pd.read_csv("tournament_data.csv").set_index("id")
feature_names = [f for f in training_data.columns if "feature" in f]

# train a model to make predictions on tournament data
model = XGBRegressor(max_depth=5, learning_rate=0.01,
                     n_estimators=2000, colsample_bytree=0.1)
model.fit(training_data[feature_names], training_data["target"])

predictions = model.predict(tournament_data[feature_names])
predictions.to_csv("predictions.csv")
```

در این بخش میخواهیم در خصوص ماهیت مسئله time series توضیح دهیم . این مسئله دامنه وسیعی از صنایع و کسب و کارها با خود مرتبط میکند و هدف کلی ما استفاده از داده های زمانی گذشته برای پیش بینی رویدادهایی در آینده می باشد . این امر به سیاست گذاران و برنامه ریزان این قابلیت را میدهد که برای آینده برنامه ریزی مناسب داشته باشند و برای آینده و رویداد های پیش رو آماده باشند . امروز یکی از کاربرد های مهم و وسیعی که در کشور و جهان در بحث سری های زمانی مشاهده میشود ، بحث بازار های مالی و پیش بینی آینده بازار های مالی میباشد . هم چنین با شیوع بیماری همه گیر کورونا در جهان تحلیل سری های زمانی تعداد بیماران یافته شده و تعداد مرگ میر میتواند درک درستی از بیماران در مناطق جغرافیایی و برنامه ریزی برای تامین نیازهای بهداشتی آن منطقه جغرافیایی به ما بدهد.
مسئله سری زمانی را میتوان به فرمت زیر فرمول بندی کرد :

فرض میکنیم y_t بیانگر مقدار یک متغیر در زمان t باشد . آنگاه یک پیش بینی برای رویدادی در آینده به صورت مثال در زمان $T+h$ با استفاده از داده های تا زمان T را میتوان به فرمت زیر نوشت .

$$\hat{y_{T+h}} = f(y_T, y_{T-1}, \cdots) \qquad (۱)$$

در فرمول (۱) ، $f(.)$ بیان گر یک تابع مناسب است که ورودی آن اطلاعات زمان گذشته و حتی زمان حال میتواند باشد .
در خیلی از موارد زمانی که ما تنها با یک سری زمانی یا اصطلاحا Univariate time series سروکار داریم ، استفاده از یک تابع خطی میتواند راهگشا باشد و مسئله ما را حل کند.

$$\hat{y_{T+h}} = v + \alpha_1 y_T + \alpha_2 y_{T-1} + \cdots$$

در مسائلی مثل مسائل اقتصادی و یا در پروژه جاری درس مقدار یک متغیر فقط وابسته به یک متغیر پیشین نمیباشد و از چندین متغیر پیشین تاثیر میگیرد .به طور مثال یک متغیری مثل مخارج و هزینه های یک خانواده وابسته به متغیرهایی مثل درآمد ، سود بانکی و سرمایه گذاری های خانواده میباشد .
این مسئله جدید را میتوان به صورت زیر فرمول بندی کرد .
متغیرهای وابسته $y_{1,t}, y_{2,t}, \cdots, y_{K,t}$ را برای پیش بینی $y_{1,T+h}$ استفاده میکنیم . پس داریم :

$$\hat{y_{1,T+h}} = f_1(y_{1,T}, y_{2,T}, \cdots, y_{K,T}, y_{1,T-1}, y_{2,T-1}, \cdots, y_{K,T-1}, y_{1,T-2}, \cdots) \qquad (۲)$$

به صورت عمومی برای k امین متغیر خواهیم داشت :

$$\hat{y_{k,T+h}} = f_k(y_{1,T}, \cdots, y_{K,T}, y_{1,T-1}, \cdots, y_{K,T-1}, \cdots, y_{K,T-1}, \cdots) \qquad (۳)$$

به مجموعه از سری های زمانی $y_{k,t}, k = 1, \cdots, K, t = 1, \cdots, T$ را یک multiple time series میگوییم .

۳ راه حل ها

بنا بر مقالات مورد بررسی در حوزه سری های زمانی راه حل ها و الگوریتم های معروفی وجود دارند که عبارتند از :

۱. AutoRegressive (AR)
۲. AutoCorrelation(AC)
۳. Moving Average(MA)
۴. AutoRegressive Integrated Moving Average (ARIMA)
۵. Convolution Neural Network(CNN)
۶. LSTM
۷. FaceBook Prophet
۸. XGBoost

۱.۳ Autoregressive

همان طور در فصل مقدمه مطرح شد یکی از روش های کارآمد در سری های زمانی استفاده از یک ترکیب خطی است . دقیقا برای مسئله چند سری زمانی هم این ایده را استفاده میکنیم .
بنا بر فرمول (۱) و برای یک سری زمانی با $h = 1$ برای پیش بینی خواهیم داشت :

$$\hat{y_{T+1}} = v + \alpha_1 y_T + \alpha_2 y_{T-1} + \cdots$$

تعداد داده های پیشین که در مسئله استفاده میکنیم را محدود میکنیم و تعداد آن را p فرض میکنیم پس برای پیش بینی خواهیم داشت :

$$\hat{y_{T+1}} = v + \alpha_1 y_T + \cdots + \alpha_p y_{T-p+1} \tag{۴}$$

همانند بسیاری از مسائل در حوزه پیش بینی ، ما مقداری واقعی را y_{T+1} در نظر میگیریم و مقدار پیش بینی $\hat{y_{T+1}}$ دارای خطایی نسبت به y_{T+1} است . خطای این پیش بینی را به صورت زیر تعریف میکنیم .

$$u_{T+1} = y_{T+1} - \hat{y_{T+1}}$$

پس خواهیم داشت :

$$y_{T+1} = \hat{y_{T+1}} + u_{T+1} = \alpha_1 y_T + \cdots + \alpha_p y_{T-p+1} + u_{T+1} \tag{۵}$$

در حالت خیلی ساده و تک سری زمانی میتوان مسئله به صورت زیر تعریف کنیم .

$$X(t+1) = \alpha_0 + \alpha_1 * X(t-1) + \alpha_2 * X(t-2)$$

۲.۳ AutoCorrelation

یک مدل autoregression با فرض ساده ای که داده های زمان آینده وابستگی به داده های قبلی ارتباط دارد و این ارتباط خطی میباشد ، جلو میرود . این ارتباط بین متغیرها را کورولیشن مینامیم .
اگر هردو متغیر در یک جهت تغییر کنند (به طور مثال افزایش یک متغیر سبب افزایش متغیر دیگر شود) این اتفاق را کورولیشن مثبت مینامیم . دقیقا مخالف این روند اگر متغیر در خلاف جهت هم واکنش نشان دهند (به طور مثال افزایش یک متغیر سبب کاهش متغیر دیگر شود) این اتفاق را کورولیشن منفی مینامیم .
کورولیشن بین دو متغیر از روش های آماری قابل محاسبه است و متغیرهایی که با هدف کورولیشن قوی تر دارند را میتوان در مدل AR ، وزن بیشتری را اختصاص داد . چون ما کورولیشن بین یک متغیر با خود متغیر در بازه های زمانی گذشته بررسی میکنیم آن را Autocorrelation می نامیم .
یکی دیگر از کاربرد های Autocorrelation این است در صورتی که خروجی کورولیشن بسیار ضعیف باشد میتوان مطرح کرد که شاید مسئله قابل پیش بینی نباشد و سراغ روش های و الگوریتم های مسائل غیر از سری زمان رفت .

۳.۳ MA

روش میانگین متحرک ارتباط تنگاتنگی با روش AR دارد و به صورت زیر تعریف میگردد .
فرض کنید قصد داریم متغیر y_T را پیش بینی کنیم . خواهیم داشت :

$$y_T = \alpha_1 y_{T-1} + \alpha_2 y_{T-2} + \cdots + \alpha_p y_{T-p+1} + u_T \quad (۶)$$

$$y_{T-1} = \alpha_1 y_{T-2} + \alpha_2 y_{T-3} + \cdots + \alpha_p y_{T-p} + u_{T-1} \quad (۷)$$

حال برای روش میانگین متحرک داریم :

$$y_T = \beta_1 u_T + \beta_2 u_{T-1} + \cdots + \beta_p u_{T-p} \quad (۸)$$

که هدف در اصل یافتن وزن ها میباشد . در این جا بتا ها میباشد . تمامی معادلات MA،AR را میتوان برای درجات مختلفی تعریف کرد که ساده ترین فرم آنها فرم درجه ۱ آنها میباشد یعنی متغیر آینده فقط وابسته به یک گام قبلی است .

۴.۳ ARIMA

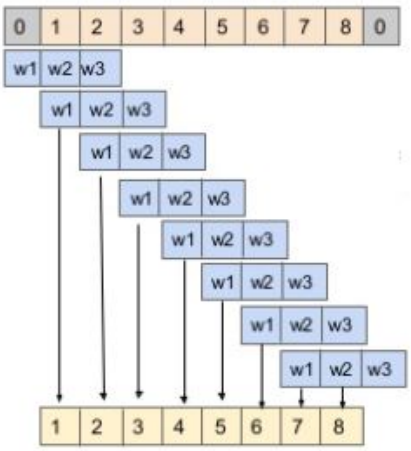
به صورت خلاصه مدل ARIMA ترکیبی از مدل MA با درجه q یعنی پنجره های ما به طول q است و مدل AR با درجه P میباشد . در این مدل ما پس از مشخص کردن مقدار مناسب q،p ما مانند دو حالت قبل به یافتن وزن ها میپردازیم .

$$y_T = \alpha_1 y_{T-1} + \alpha_2 y_{T-2} + \cdots + \alpha_p y_{T-p} + \beta_1 u_T + \beta_2 u_{T-1} + \cdots + \beta_q u_{T-q} + c \quad (۹)$$

در فرمول فوق c یک ثابت میباشد .

۵.۳ CNN

شبکه های کانولوشنی یک بعدی را میتوان در تحلیل و تفسیر سری های زمانی تک متغیره استفاده کرد . هم چنین در صورتی که با سری های زمانی چند متغیره سرکار داشته باشیم میتوانیم با استفاده از چند جریان موازی آنها را مدیریت کنیم چرا این جریان دارند به صورت همزمان رخ میدهند .
به صورت کلی در عمل کانولوشن یک بعد پنجره هایی در طول زمان بر روی داده ها اعمال میکند . عملیات کانولوشنی به نوعی مکان داده را حذف میکند و ما مکان داده را خواهیم دانست که در این چون یک بعد در امتداد زمان داریم ما زمان داده را پس از اعمال تغییر خواهیم داشت .
مزیت این شبکه ها این است که میتوان داده خام را به صورت مستقیم به آنها داد و نیازی به بحث های مهندسی ویژگی وجود ندارد .
در اصل بردار وزن ها به طول پنجره مشخص بر روی داده ها عمل کانولوشن انجام میدهند و خروجی میگیرند .



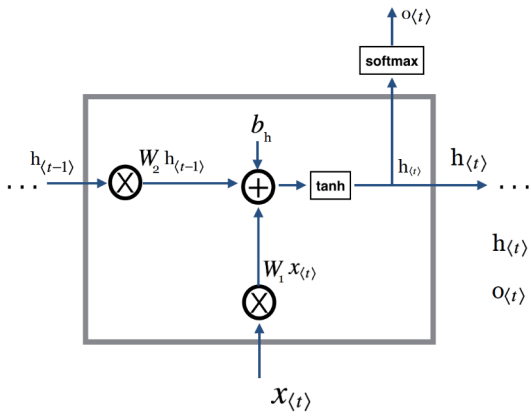
شکل ۲: مثالی از کانولوشنی یک بعدی

۶.۳ LSTM

شبکه های باحافظه مانند LSTM ما را قادر میکنند که بتوانیم اطلاعات زمان های قبلی را نیز نگه داری و در پردازش های زمان های بعدی از آنها بهره ببریم .
برخلاف شبکه عصبی بازگشتی سنتی که صرفا جمع متوازن سیگنالهای ورودی را محاسبه کرده و سپس از یک تابع فعالسازی عبور میدهد هر واحد LSTM از یک حافظه C_t در زمان t بهره میبرد. خروجی h_t و یا فعالسازی واحد LSTM بصورت $h_t = \Gamma_o.tanh(C_t)$ است که در آن Γ_o دروازه خروجی است که کنترل کننده میزان محتوایی است که از طریق حافظه ارائه میشود. دروازه خروجی از طریق عبارت $\Gamma_o = \sigma(W_o.[h_{t-1}, X_t] + b_o)$ محاسبه میشود که در آن σتابع فعالسازی سیگموید است. W_o نیز یک ماتریس اوریب است. سلول حافظه C_tنیز با فراموشی نسبی حافظه فعلی و اضافه کردن محتوای حافظه جدید بصورت $\hat{C}_t$ بصورت $C_t = \Gamma_f.C_{t-1} + \Gamma_u.\hat{C}_t$ بروز رسانی میشود که در آن محتوای حافظه جدید از طریق عبارت $\hat{C}_t = tanh(W_C.[h_{t-1}, X_t] + b_c)$ بدست می آید. آن میزان از حافظه فعلی که باید فراموش شود توسط دروازه فراموشی Γ_f کنترل میشود و آن میزانی از محتوای حافظه جدید که باید به سلول حافظه اضافه شود توسط دروازه بروزرسانی (یا بعضا به دروازه ورودی معروف است) انجام میگیرد. این عمل با محاسبات زیر صورت میگیرد :

$$\Gamma_f = \sigma(W_f.[h_{t-1}, X_t] + b_f) \tag{۱۰}$$

$$\Gamma_u = \sigma(W_u.[h_{t-1}, X_t] + b_u) \tag{۱۱}$$



$$h_{\langle t\rangle} = \tanh(W_1 x_{\langle t\rangle} + W_2\, h_{\langle t-1\rangle} + b_h)$$

$$o_{\langle t\rangle} = soft\max(W_3\, h_{\langle t\rangle} + b_o)$$

شکل ۳: LSTM