WIND SPEED PREDICTION

Dexin Liu, Wenjing Wang, Hao Jin

5th December, 2017 — 11th January, 2018

Introduction

随着风能重要性逐渐增加，以及风机数量逐渐增多，风速预测成为了越来越重要的一项挑战。如公式(a)所示，基本的风速预测是根据已有时间点的真实风速值，对未来时间点的真实风速进行预测。然而本次大作业，我们的数据来源更加丰富，除了各个时间采样点的真实风速预测值以外，我们还有对于各个时间点真实风速的初步估计。换言之，本次大作业之中风速预测的模型可以由公式(b)来表示。

$$\{w_1, w_2, \dots, w_t\} \rightarrow \{\widetilde{w_{t+1}}, \widetilde{w_{t+2}}, \dots, \widetilde{w_{t+i}}\}$$

(a)

$$\begin{matrix} \{w_1, w_2, \dots, w_t\} \\ \{p_1, p_2, \dots, p_t\} \end{matrix} \begin{matrix} \searrow \\ \nearrow \end{matrix} \{\widetilde{w_{t+1}}, \widetilde{w_{t+2}}, \dots, \widetilde{w_{t+i}}\}$$

(b)

方法层面上，新的问题要求需要我们对传统风速预测模型进行改进。我们提前获得的对未来风速的预测，在一定程度上反映了真实风速的变化趋势，但是这些预测值与对应点的真实值又存在着并不可以忽略的误差。基于这样的感性认识，不难发现传统风速预测方法在本次大作业的设定下存在着明显的缺陷：如果不采用预先获知的预测风速，预测模型就没有充分使用数据集提供的信息；如果盲目的将预先获知的预测风速加入到传统模型之中，实测数据与预先获知的预测风速之间的矛盾会负面影响预测模型的表现结果。

数据层面上，需要采取多种方法对原始数据进行清洗与整理，才能真正进行预测模型的训练。原始数据中与风速预测密切的主要有两部分：实测风速点的真实风速，预测风速点的预测风速。原始数据的噪音主要来自于真实风速中的自我矛盾，在这里指的是有个别实际观测点给出了两个实际风速测量值。由于这一类错误只出现在两个时间点上，我们直接删除了互相矛盾的两个实际风速测量值中的一个，消解了矛盾。原始数据的不统一体现在机器个体时间间隔不统一，不同机器时间间隔不统一，预测风速点与实测观测点之间时间间隔不统一等方面。具体而言，这里的不统一指的是相邻的两个时间点的间隔不一致，由此引入的变量会为模型训练增加不必要的难度。因此需要合理的方法，对这些数据点之间间隔的不统一进行标准化。

模型的训练层面上，需要对原始问题空间进行重新整合。换言之，由于原始数据中真实风速序列，预测风速序列都在时间维度上存在不同程度上的缺失，所以可供进行模型训练的连续时间点并不是全体数据点。因此，根据具体的数据缺失情况，需要重新定义问题，相应地划分训练集与测试集。

Problem Definition and Data Preprocessing

I. Problem definition

我们对每台机器在整十五分钟点的实际风速进行预测。

具体来说，对于给定的时间点集合 $\{t_1, t_2, .., t_k, t_{k+1}, .., t_n\}$，其中 $t_{i+1} - t_i = 15min, \forall 1 \leq i \leq n-1$。已知的数据包括两部分：部分已知的实测风速数据 $\{w_1, w_2, .., w_k\}$，所有数据的预先得知的预测风速$\{p_1, p_2, .., p_n\}$。以上式子中w_i，p_j分别表示的是在第t_i，t_j时刻的实际风速和预测风速。

以上问题的定义在实际应用中是合理的，并且可以从原始数据中构建这样的问题：

现实应用中，一定存在可以将已知实测风速的时间点与未知实测风速的时间点分割的时间点 t_k。在原先的题目要求下“可以使用的数据有：截至当天（应该是前一天）12:00:00之前的历史测量风速；…”，意味着对于任意一个时间点所在的时间点集合，t_k 这一时间点就应该是前一天 12:00 的时刻。除此之外，由于预先获得的天气预报的预测区间总是长于 24h，对于任意需要预测实际风速的时间点，我们一定可以从这一天之前的天气预报中获得对这一天实际风速的若干个预测风速值。

原始数据提供了秒级实际测量风速，与每一天对未来三天的预测风速。其中，实际测量风速的时间间隔并不统一，但是都一致地不超过一分钟；预测风速的时间点间隔本身就是十五分钟，是满足我们问题的定义的。然而，由于预测风速都是从当天的 2:00 开始到第三天的 24:00，这就导致了对于相同的时间点存在多个不同的预测值。即使原始数据本身存在或多或少的不统一，然而不难发现我们定义的问题的要求实际上低于原始数据的复杂度，这就意味着本问题是可以从原始数据中构架出来的。

II. Data preprocessing

II.i 错误数据处理

数据的错误主要出现在实测风速之中。我们将错误的原始数据的截图如下展示：

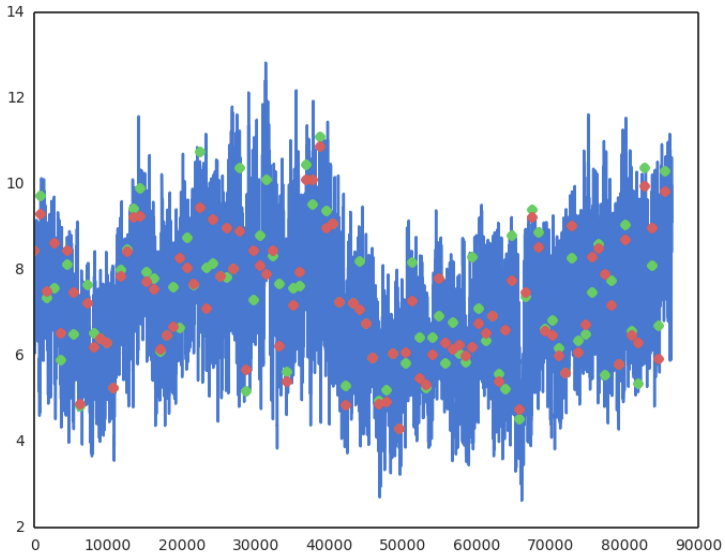
456	4196	3.0698
457	4198	2.0683
458	4201	2.2211
459	4214	1.9989
460	4214	2.6237
461	4224	3.0976
462	4232	3.3475

1107	4670	5.0694
1108	4676	4.4307
1109	4681	5.0416
1110	4684	4.6667
1111	4684	5.9026
1112	4690	4.6667
1113	4697	4.5140

错误数据分别出现在第三台风机7月2日的第4214秒的实测数据，和第四台机器7月7日的第4684秒的实测数据。这两处数据错误会导致接下来对实测数据的处理出现错误。作为解决方案，我们将同一秒数第二个数据点删除，保证了实测数据本身不出现错误。

II.ii 实测数据标准化

由问题本身的设定，我们需要获得整15分钟的各台机器的实际测量风速。由于原始数据中实际测量风速的时间间隔均小于30秒，所以我们考虑采用插值的方法进行对整十五分钟风速的估计。如下图，我们展示了在5月20日这一天对整15分钟实际测量风速插值获得的两种效果：



本张示意图横轴表示的是时间，纵轴表示的是风速大小。我们采用的第一种插值方法为取整十五分钟时间点 t_0 先后两个最相邻时间点 t_1, t_2 处风速的加权平均值，即

$$w_{t_0} = (w_{t_1} * \frac{1}{t_0 - t_1} + w_{t_2} * \frac{1}{t_2 - t_0}) \Big/ (\frac{1}{t_0 - t_1} + \frac{1}{t_2 - t_0}), t_1 < t_0 < t_2$$

特别地，如果整十五分钟时间点属于原始数据实测风速点集合，那么就把那个数据点作为整十五分钟点的实测风速。本张示意图中，红色的点表示的就是通过这种插值方法获得的整十五分钟实测风速点。我们采用的第二种插值方法为三次样条法，即使用分段三次函数拟合已有的数据点，再在需要的时间点处取得函数值。本张示意图中，绿色的点表示的就是通过这种插值方法获得的整十五分钟实测风速点。

参考上面的可视化结果，在没有真实十五分钟实测风速点的情况下，难以比较两种方法的优良性。从直观来看，任何一种方法作为对整十五分钟时间点的实测风速的标准化都是合理的。然而相比于三次样条法，第一种插值方法计算复杂度较小，故在接下来的试验中，我们统一采取第一种方法获取整十五分钟时间点处的实测风速。

II.iii 时间点标签化

每一个观测点所处的小时、日期、月份会否与该位置的实际测量风速是否有关联，这一问题将影响数据中所有点在训练中是否是对等的。换言之，如果它们之间存在关联，我们

的模型训练必须考虑到他们之间的规律性，也就是说需要把时间也作为变量考虑到模型之中。如果它们之间不存在关联，我们就可以把每一个时间点认为是无差别的时间点。

	Month	Date	Hour	Minute	Speed
Month	1.000000	-0.544288	-0.100450	-0.025034	-0.000207
Date	-0.544288	1.000000	-0.251692	-0.062738	-0.000512
Hour	-0.100450	-0.251692	1.000000	-0.282916	-0.001614
Minute	-0.025034	-0.062738	-0.282916	1.000000	0.008564
Speed	-0.000207	-0.000512	-0.001614	0.008564	1.000000

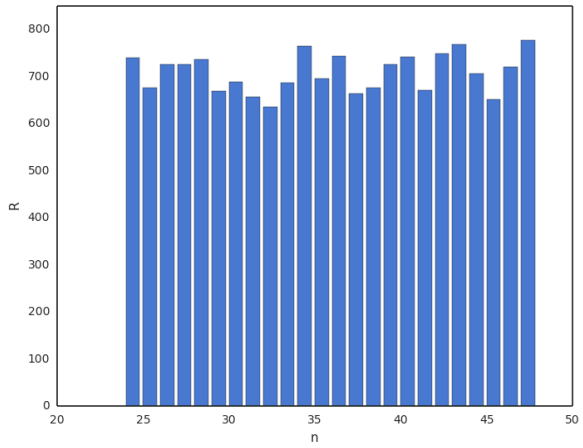
针对这个问题，我们采用了如下两种方法进行检验：

首先我们直接计算原始数据之中实测风速，月份，日期，小时之间的相关系数。特别地，针对第一台风机根据II.ii得到的实测风速点，我们可以得到实测风速与月份，日期，小时，分钟的相关系数。风速与时间的相关系数都小于0.1.其他的风机都能够得出类似的结果，在这里不再赘述。由此，我们可以认为时间与风速的关联性十分小，在建立模型的时候，可以将所有时间点认为是没有区别的。

其次，风速变化与时间的相关性等同于风速变化是否在时间序列上存在规律。因此，我们任意取连续八个小时的时间点的实际测量风速序列$x = \{w_t, w_{t+1}, \ldots, w_{t+31}\}$，然后统计该风速序列与 $n(24 \leq n < 47)$ 小时之前实际测量风速序列$y = \{w_{t-4n}, w_{t+1-4n}, \ldots, w_{t+31-4n}\}$之间的相似度。在这里我们采用的是信号分析中经常采用的互相关函数：

$$R_{xy}(t) = \frac{1}{32}\sum_{i=0}^{31} x(i) * y(i+t)$$

其中 t 表示循环位移。对于第一台风机，我们可以获得如下数据：



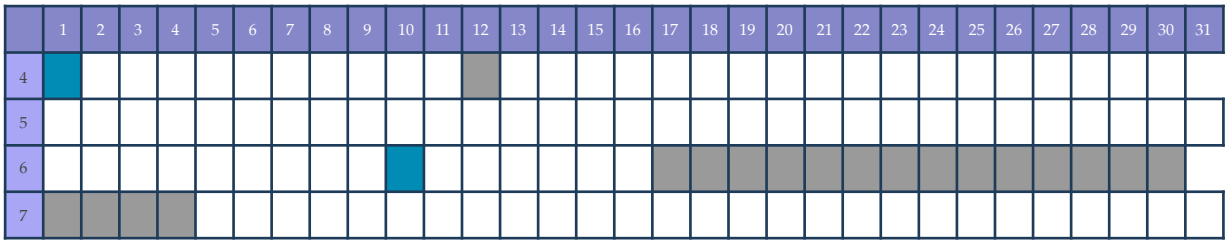
我们可以发现时间间隔的改变并没有影响两个风速序列 x, y 的相关性，可以从另一个侧面说明时间与实测风速关系不大，所有时间点可以平等对待。

II.iv 预测数据的降维

预测数据在两个尺度上具有多维度的特性。首先，同一个时间点可能存在多个不同时间对其的预测值。其次，一个预测值实际上是一个长度为七的向量。在一些模型的输入中，需要对预测数据进行降维，这方面的展开将在模型之中详细介绍。

II.v 测试集与训练集的划分

预测数据与实测数据存在不同程度上的缺失，导致原始数据整体并不能构成一个连续的时间点序列。我们将预测数据与实测数据的确实情况展示如下：



图中显示了天气预报以及实测数据的缺失情况，图中灰色的日期在六台风机的实际测量风速上不同程度上存在缺失，蓝色的日期在 0:00 - 2:00 之内缺失天气预报。

考虑到原始数据的缺失情况，我们选定 4.13 - 6.9 为训练集，7.5 - 7.31 为测试集。

Model Construction

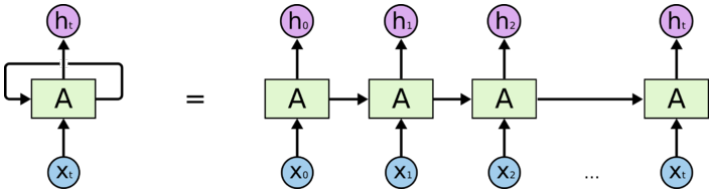
风速预测是一种基于时间序列的预测，可以采用传统的回归方法，也可以采用神经网络模型。我们的模型对两种方法都有所使用，在以下部分中将分别介绍模型之中的神经网络模型以及回归方法。

I. 神经网络模型

LSTM是一种特别的RNN(循环神经网络)，是分析时间序列型数据的常用网络模型。风速预测的一种传统方法也是通过LSTM对未来的时间点进行预测。为了解决我们提出的预测问题，我们也采用了LSTM网络作为我们模型的一部分。在以下部分，我们先对LSTM进行简要介绍，再详细说明基于我们的预测问题，我们对LSTM网络做的改进。

I.i LSTM

循环神经网络(RNN)本质上是同一神经网络的多次复制，每个神经网络模块会把信息传递给下一个模块。链式的特征使其成为了能够处理序列类数据的一种自然的网络结构:



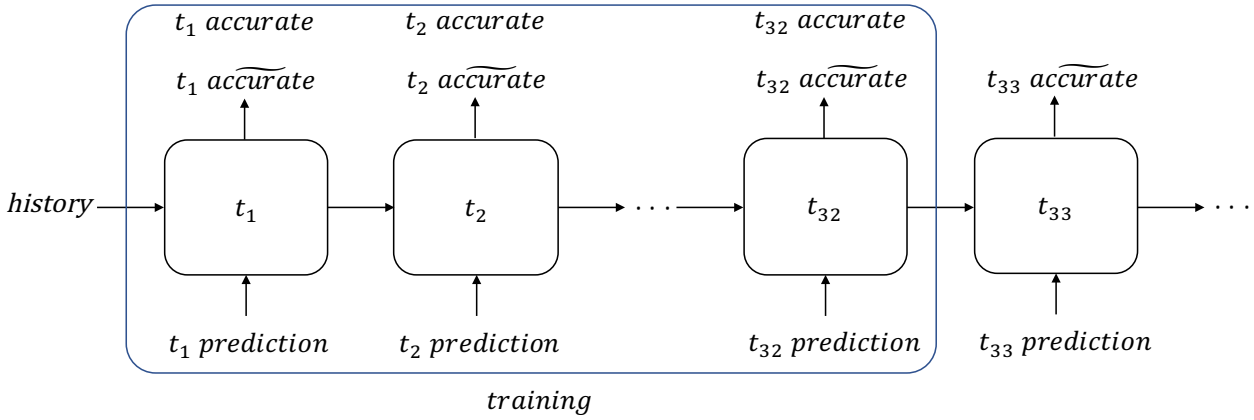
LSTM 是一种特殊的循环神经网络，相比于一般的循环神经网络，LSTM 重复模块A之中的结构更加复杂。

除了一般的循环神经网络输入门之中的 tanh 层，LSTM模块之中还包括 sigmoid 层构成的遗忘门用以更新细胞状态，以及由 sigmoid 层和tanh层构成的输出门用以提供结合了细胞状态的下一个模块的历史数据输入。通俗来讲，LSTM 会根据之前的输入信息和输出信息对现有的输入信息进行处理，其中过去信息的影响作用会随着距离现有单元的距离的增加而减少。

I.ii 用于风速预测的LSTM模型

不失一般性，预测问题中时间点集合$\{t_1, t_2, \ldots, t_k, t_{k+1}, \ldots, t_n\}$中 $t = 32$。换言之，我们根据已有的 32 个时间点的实际测量风速，以及所有时间点能获取到的合理的天气预报值对

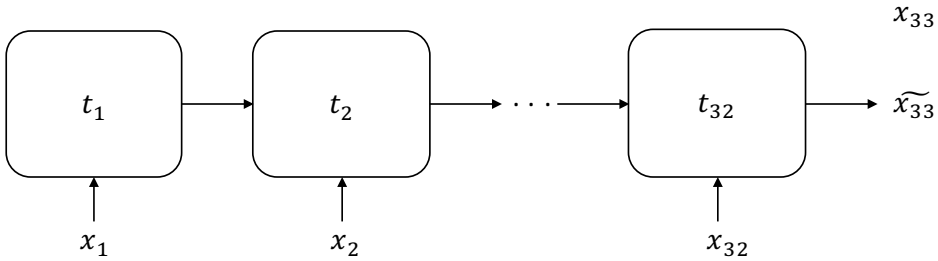
未来任意长度的时间点段处的实际风速进行预测。使用 LSTM 模型对未来时间点的实际风速进行预测的话，我们提出了两个 LSTM 模型：



第一个 LSTM 模型如上图所示，这是一个序列到序列的网络。在给定实测风速的32个时间点上进行模型的训练，每一个 LSTM 模块综合现有的对当日的天气预报以及历史数据对一个时间点进行实测风速的预测。模型在连续的给定实测风速的模块上根据损失函数完成训练，其中损失函数如下公式所示：

$$Loss = \sum_{i=1}^{32} (t_i\ \widetilde{accurate}\ - t_i\ accurate)^2$$

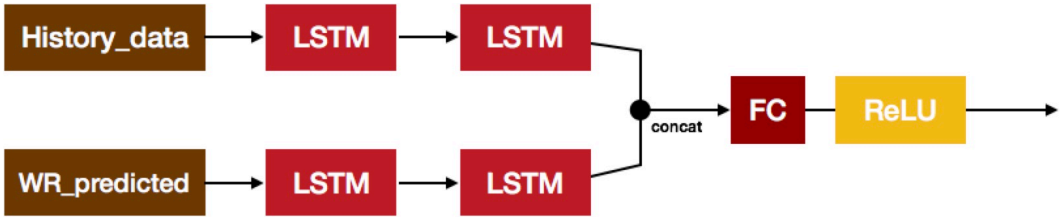
这个 LSTM 比较符合风速预测问题的设定，通过给定实测风速的时间点序列 $\{t_1, t_2, .., t_{32}\}$ 上进行训练，达到对未来时间点序列 $\{t_{33}, t_{34}, ..\}$ 进行预测。模型之中，实际测量风速通过训练集上的损失函数参与训练，每一天的天气预报作为预测的输入。



第二个 LSTM 模型的基础单元如上图所示，这是一个序列到点的预测方案。模型通过接手前面的 32 个时间点的某种信息，对第 33 个时间点上可能并非同一种类的信息进行预测，这样产生的损失函数如下：

$$Loss = (x_{33} - \widetilde{x_{33}})^2$$

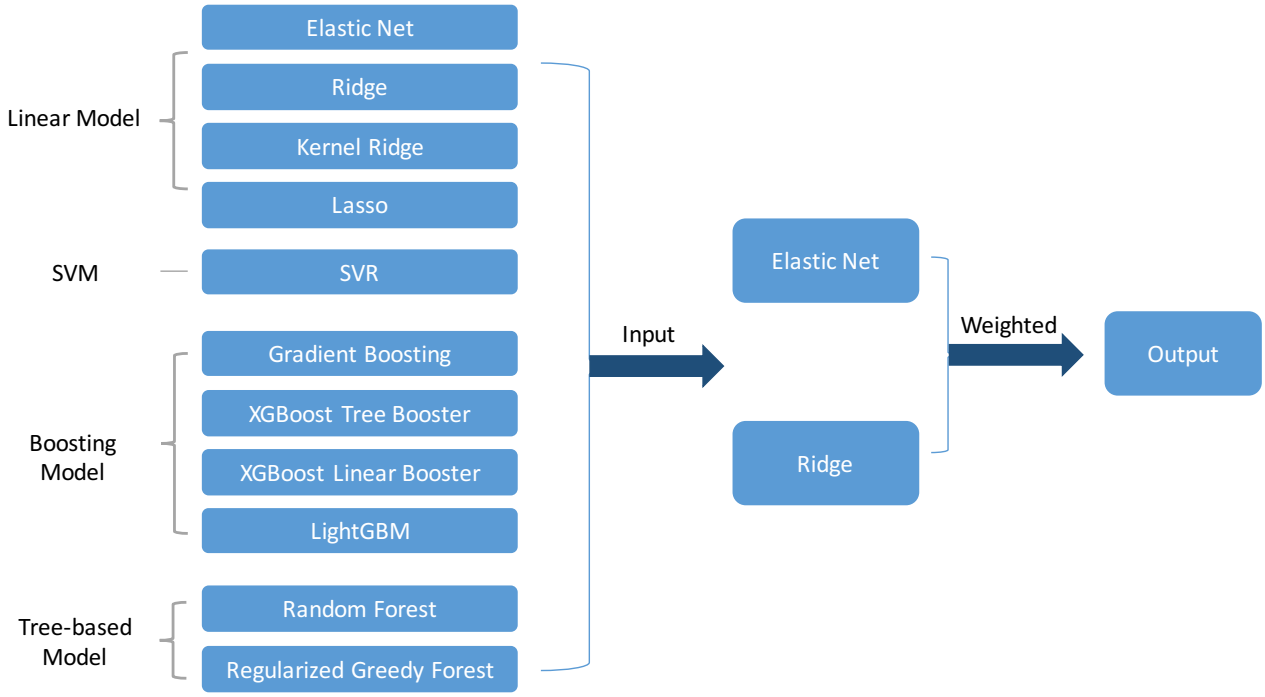
损失函数经过回传整个循环网络，完成对模型的训练。虽然该模型的基础单元是一个序列到点，但是仍然有能力完成序列到序列的预测任务。只要将预测序列向后移动一个时间点的距离，将上一轮的预测值视为该时间点的准确值，即可完成下一时间点的预测任务。我们并没有明确每一个 LSTM 模块的输入信息，也就意味着这个 LSTM 模型既可以根据历史实测数据对未来实测数据进行预测，也可以根据天气预报对未来实测数据进行预测。按照这样的思路，我们构造了下图结构的预测网络：



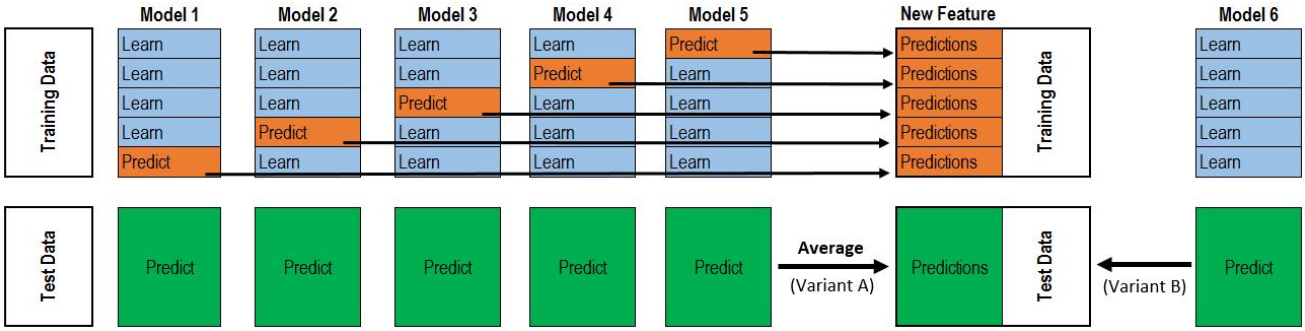
本网络结构图中 LSTM 表示的就是如上介绍的 LSTM 序列到点的基础单元。天气预报数据，历史数据均作为输入进入模型之中，既不损失原始数据集中的信息，又在 LSTM 输入的层面分离历史数据，天气预报。

II. 回归方法

单一回归模型越发难以适用于很多预测问题的要求，需要采用模型融合的方法。



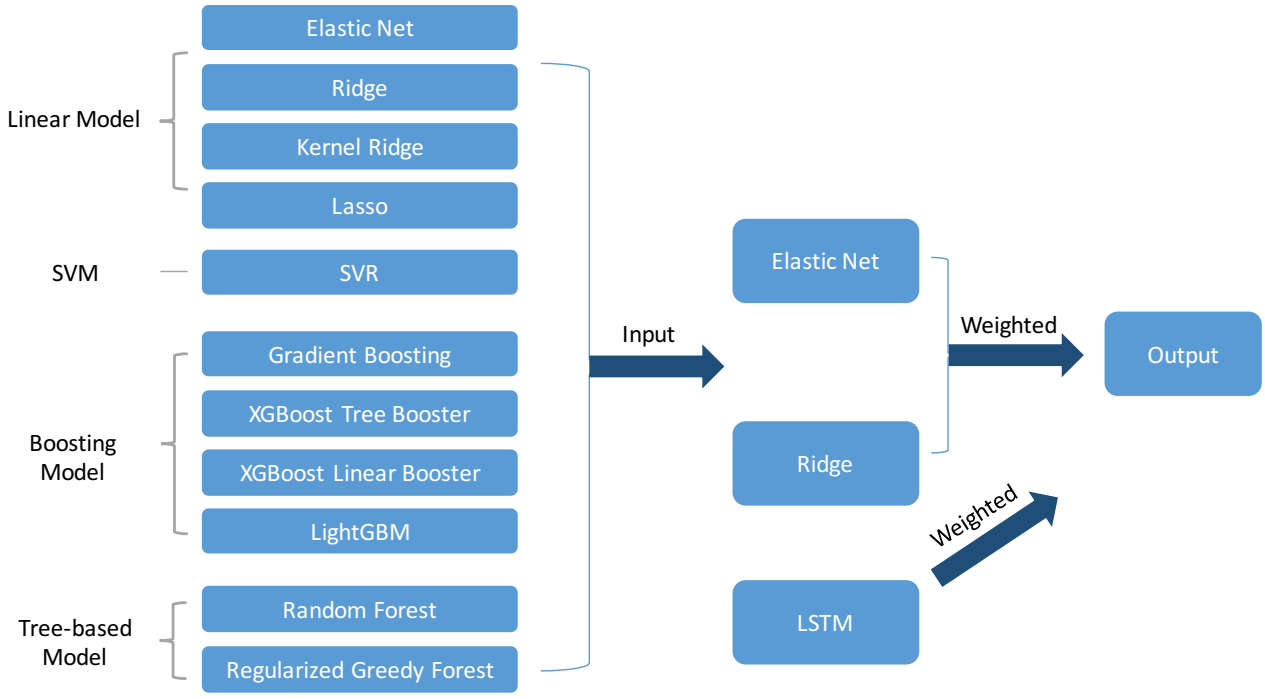
基础模型包括上图所示的四类模型：支持向量机(SVM)、Boosting模型、线性模型、树形模型。这些基础模型通过贝叶斯优化(Bayesian Optimization)的方法在训练集上确定了最佳参数。



在完成了对基础模型的训练之后，我们选择使用一个三层Stacking模型组合各个模型的结果。整个Stacking的过程类似于CV验证：将训练集分为五份，对每个基本模型进行5轮训练，一次使用其中的4份作为训练集训练，预测余下一份的结果，5轮后得到训练集大小的预测数据；同时在每轮中对测试集进行预测，对每个基本模型来说测试集的预测结果为5轮结果的均值；在第二层中，我们输入(训练集上的预测结果*基本模型数量)的数据进行训练；第三层最终输出为第二层的预测结果。

III. LSTM + 回归模型

我们将LSTM部分的模型作为stack模型的一部分，添加到其中，可以获得如下图的最终模型网络：



Experimental Result

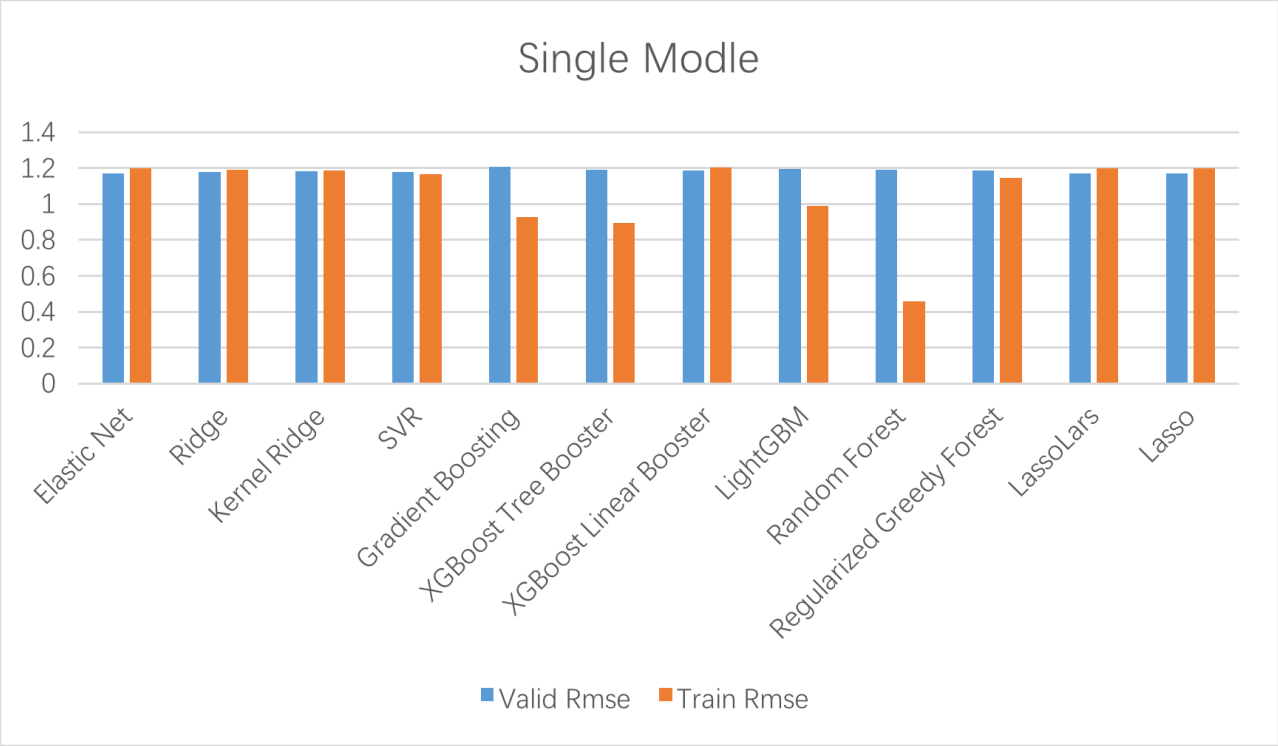
I. 评价标准

为了定义风速预测模型的优良性，我们如下定义测试集上的误差函数：

II. 模型训练以及交叉检测

II.i Stack 模型训练

首先使用对Stacking模型中的各个基本模型进行参数选择：使用Grid Search和Randomized Search方法进行调参，组成Stacking各个部分的预测调整到接近的部分。最终我们得到的基本模型预测结果的RMSE都接近于1.2，这样接近的结果对我们接下来模型融合十分有利。



之后我们需要对Stacking模型进行训练。需要注意的是，之前在整个训练集上训练得到的基本模型已经不能使用，因Stacking中需要对分为5折验证的数据分别训练一次。此外Stacking中的融合模型也是十分重要的，这会直接影响各个模型间能否互相补充、检验其他模型中的异常值。在回归问题的Stacking模型中，常常选用简单的线性函数作为stacker输出最后的结果。经过权衡后我们选择了ElasticNet和Ridge两种stacker模型，权重为[0.6, 0.4]，调参之后这两个模型已经可以很好的提升模型融合的效果。

II.ii LSTM 模型训练

上文提到我们在LSTM部分选择了一个上部为两个分别接受历史数据和历史预测数据输入的双侧LSTM，下部使用多层全连接层降维的LSTM模型。这一模型的时序层较多，即内部权重参数众多，训练较为耗时。由于LSTM本身的特性，相对而言这一模型训练需要的epoch数是较少的，几乎在20次训练之内就可以达到收敛的Validation RMSE低值。
基于以上我们LSTM模型的特性，在时序神经网络训练的过程中我们采取了多种方式防止训练过度（过拟合）：

-Alterable learning rate： 训练过程中逐渐减小训练速度，在训练过半后LR即将为初始的25%

-Early stopping：记录每次validation的测试值，在RMSE有上升趋势时停止训练，并且将模型回退到RMSE处于低谷时的阶段

II.iii 模型验证与结论

最终我们得到Staking模型与LSTM模型按照[0.7, 0.3]加权求得的结果，在测试集上测得的RMSE为1.1103296809。

虽然在这个问题下很难验证模型是否能准确地预测需要的风速值，我们还是可以看到在模型的综合与改进过程中误差在不断的下降：经过Stacking，RMSE较基本模型大约下降了10%，经过LSTM的参数调整和结构优化，它的预测准确度也较之前有了约10%的提升。在于其他组同学交流的过程中，发现我们采用的方法可以说是比较全面的，兼顾了序列预测常用的时序神经网络方法以及较为高效的统计回归方式，并且采取了合理的模型融合方案。结合对问题的重新解析（十五分钟为间隔的预测），从数据处理到最后的结果预测都是比较合乎现实的，我们认为通过本项目提出的模型进行十五分钟级的风速预测是有实用价值的。