Motivación

Dentro del área de aprendizaje de maquina el aprendizaje supervisado es una técnica que consiste en aprender una función a partir de un conjunto de datos de entrenamiento. El conjunto de entrenamiento consiste en pares (datos de entrada(X), salida esperada(Y)). Utilizando estos datos se aprenderá una función que luego será capaz de predecir la salida esperada utilizando datos de entrada nunca vistos. Cuando la salida de la función aprendida es un valor continuo la tarea de aprendizaje se denomina regresión.

El análisis de regresión es un conjunto de técnicas de modelado y análisis numérico la cual consiste en determinar el valor de una variable dependiente a partir de un conjunto de variables independientes. Esta técnica es utilizada para realizar pronósticos, inferencia, testeo de hipótesis y evaluación de causalidad entre otras cosas.

Dentro del aprendizaje supervisado existen diferentes modelos de regresión los cuales son más o menos eficientes de acuerdo a la naturaleza de los datos del problema en cuestión. Ejemplos de ellos son: Regresión Lineal[1], Maquinas de Soporte Vectorial[2] , Redes Neuronales[2], Arboles de decisión[3]. Estos algoritmos son utilizados en las áreas de agricultura, finanzas, historia, industria, relaciones laborales y ciencias del medioambiente, entre otras.

Nuestra propuesta consiste en la evaluación de un conjunto de algoritmos de regresión en el marco de un caso de estudio específico. A continuación describiremos el mismo.

Caso de estudio

En la actualidad existen una variedad de deportes acuáticos que se ven beneficiados por los pronósticos acerca del oleaje en una zona cercana a la costa. Ejemplos de estos pueden ser el surf, windsurf, kitesurf, kayac… Los deportistas están sumamente atentos a las condiciones que presenta el mar para evaluar la playa en la que van a practicar el deporte. Utilizando un pronóstico de oleaje el deportista puede planificar sus actividades para lograr estar presente en el lugar con mejores condiciones. Eventualmente el deportista puede evitarse recorridas extensas buscando el sitio con mejores condiciones del día.

Actualmente los pronósticos de oleaje son derivados a partir de la resolución de un modelo matemático global sumamente complejo. Este provee de la predicción de la altura de las olas la dirección de las mismas y el tiempo entre una ola y otra. Si bien este pronóstico tiene buen rendimiento en alta mar, no sucede lo mismo en la cercanía de la costa. Esto se debe a que la interacción de las olas con el lecho marino cercano no es contemplado por estos modelos. De acuerdo a la geografía de la costa puede suceder que en un rango de 15 km un lugar presente condiciones apropiadas para la práctica del deporte mientras que en otro las condiciones sean desfavorables. Por esta razón es que la utilización de un pronóstico preciso en las cercanías de la costa es de suma utilidad. Existen modelos para contemplar el comportamiento de las olas en las cercanías de la costa. Los mismos reciben como entrada los datos de pronóstico de alta mar e información acerca del lecho marino. Utilizando estos datos los mismos proveen un refinamiento al pronóstico inicial. Si bien estos modelos poseen un buen comportamiento, su limitación esta relacionada con la obtención de los datos del lecho marino. Usualmente la obtención de estos datos es costosa y no se encuentra disponible en todos los lugares.

Nuestra propuesta consiste en utilizar modelos de aprendizaje de máquina para aprender de casos pasados como un determinado pronóstico de alta mar se vio reflejado en las cercanías de la costa. De esta manera ante un nuevo pronostico de alta mar se podrá predecir cual será el efecto del mismo en las cercanías de la costa. Nos centraremos en la predicción de la altura de la ola en la cercanía de la costa, mejorando la predicción del pronóstico de alta mar. Para realizar esta predicción utilizaremos métodos de regresión los cuales tendrán como datos de entrada: el pronóstico de alta mar y la observación de la altura en la cercanía de la costa.

Existen distintos trabajos que enfocan el tema de predicción de oleaje [4,5]. En ambos se utilizan datos de alta mar junto con observaciones visuales, relevadas por un observador experto, de la altura de las olas en las proximidades de la costa. Con estos datos se entrena una red neuronal que pueda estima la altura de la ola en la cercanía de la costa. En [4] los datos pertenecen a la costa australiana. Se entrena una red neuronal del tipo feed forward y se realiza una comparación entre un método numérico para refinamiento del modelo de alta mar y un método de redes neuronales, dando este ultimo una mejor precisión en los pronósticos. En [5] los datos son de la costa norte de Irlanda. Se experimenta con un ensamble de redes neuronales del tipo MDN(mixture density networks) y la salida de la red neuronal representa una distribución de probabilidad de la altura de la ola en la cercanía de la costa. Los resultados encontrados fueron positivos haciendo del modelo de red neuronal desarrollado útil para la estimación del tamaño de la ola en la cercanía de la costa.

**Referencias:**

[1] Data Mining: Practical Machine Learning Tools and Techniques, Second Edition, Ian H.Witten y Eibe Frank – Chapter 4 Linear Models.

[2]Data Mining: Practical Machine Learning Tools and Techniques, Second Edition, Ian H.Witten y Eibe Frank – Chapter 6 Extending Linear Models.

[3]Data Mining: Practical Machine Learning Tools and Techniques, Second Edition, Ian H.Witten y Eibe Frank – Chapter 6 Decision Trees

[4] Empirical Estimation of Nearshore Waves From a Global Deep-Water Wave Model Browne, M. Strauss, D. Castelle, B. Blumenstein, M. Tomlinson, R. Lane, C. Centre for Coastal Manage., Griffith Univ., Brisbane, Qld.;

[5] Predicting probability distributions for surf height using an ensemble of mixture density networks, Michael Carney, P´adraig Cunningham, Jim Dowling and Ciaran Lee, ACM International Conference Proceeding Series; Vol. 119.