Transporte Optimo

August 19, 2017

1 TGP: Transport Gaussian Processes

El fin es construir procesos estocásticos no paramétricos que se puedan utilizar como modelos bayesianos predictivos para contextos de machine learning, que sean más expresivos que los procesos guassianos (GP) pero que conserven las buenas propiedades para su entrenamiento y predicción. La idea es utilizar las herramientas de transporte óptimo para nuestro fin de la siguiente forma: a partir de un GP de referencia, deseamos encontrar (entrenar) una transformación (transporte) que transporte el GP al proceso estocástico target, minimizando algún costo ad-hoc al problema. En papers anteriores se han utilizado transportes en el contexto de inferencia bayesiana, pero siempre donde la dimensión es fija y no incluyen un costo de transporte. A continuación comentaré cada página de mis notas:

1. El problema de transporte original lo reescribo como una versión regularizada, donde la restricción de la marginal target se suma al funcional a optimizar, siguiendo la idea de los multiplicadores de Lagrange. Para comparar las distribuciones se puede utilizar cualquier divergencia positiva, con tal que cumple que $D(p,q) = 0 \Leftrightarrow p = q$. En nuestro caso utilizamos la divergencia de Kullback–Leibler $D_{KL}(p,q) = \int p \ln \frac{p}{q}$, pero hay 4 formas diferentes de utilizarla (por su asimetría y la invertibilidad del transporte).

2. En nuestro contexto de predicción, la distribución objetivo es la distribución predictiva real, la cual no tenemos acceso a evaluarla, pero si tenemos acceso a muestras (datos reales). Esto nos permite realizar una aproximación de monte carlo a una de las integrales de KL, mientras que el otro término es irrelevante para el problema de optimización. El costo de transporte también puede ser aproximado utilizando estas muestras, obteniendo así un funcional explícito a optimizar.

3. Como el multiplicador de Lagrange es siempre positivo (cuando tiene a infinito, el problema regularizado converge al problema original), es posible traspasar la constante al término del costo de transporte, reinterpretando este término como el regularizador del problema. En la siguiente parte estoy interesado en construir transportes no deterministicos que, además del GP de referencia x tiene acceso a una fuente aleatoria independiente α, de modo que la distribución de y y en costo se debe marginalizar (integrar) la variable α. Un ejemplo es cuando multiplico una gaussiana por la raiz cuadrada de una gamma inversa, el proceso resultante es student-t.

4. Definiendo notación de todos los elementos, donde T es un transporte y S es su inversa. Extendemos esta notación para los transportes no deterministas.

5. Podemos notar que podría darse el caso que x dependa de α o viceversa, es decir el caso que las fuentes no son independientes. Se revisan las fórmulas de la función de probabilidades y su densidad para el caso independiente.

6. Se plantea el problema de transporte óptimo en nuestro contexto, es decir en términos de x y de α.

7. Escribo el problema de transporte optimo de forma general, y describo los elementos en el contexto de GPs: la referencia x es gaussiana, de la objetivo y tengo acceso a n muestras, puedo evaluar y generar muestras de α de forma fácil, y la evaluación de T y S es fácil.

8. El costo de transporte puede escribirse en términos de x o de y, por lo que se explicitan las 4 aproximaciones de monte carlo, según si se puede integrar con respecto a α de forma explícita o no, y si se generar muestras desde x o se utilizan las observaciones y

9. Se muestran las 4 formas diferentes de descomponer la KL, mostrando que en 3 casos es necesario evaluar la densidad objetiva, mientras que en el primer caso no es necesario evaluar ya que ese término es constante en el problema de optimización.

10. Se escribe el problema de transporte regularizado por entropía, término que se aproxima por monte carlo, y utilizando Jensen se entrega una cota inferior más simple de evaluar numéricamente.

11. Muestro que el caso trivial, con costo cero, transporte identidad y determinista, el funcional a optimizar es exactamente el mismo que en el caso GP standard, que corresponde a la negative log-likelihood (NLL). Luego extiendo el caso con una transformación no lineal aplicada a todas las coordenadas, y el funcional es exactamente el caso NLL de warped GP. En el tercer caso se toma un transporte lineal $T = \sqrt{(}\alpha)x$ con α una distribución gamma inversa, entonces el funcional obtenido coincide con el caso NLL de Student-t process.

12. En el cuarto caso, tomo que mi medida de referencia es un proceso gaussiano de ruido blanco, y construyo el transporte T a partir de un kernel de covarianza, de modo que el funcional es el mismo que el GP standard, pero con la diferencia que la medida de referencia esta fija. En el quinto caso tomamos como función de costo de transporte $C(x, y)$ a la norma. Mostramos la diferencia entre considerar una distribución de referencia fija (costo positivo) versus una entrenable con el kernel (costo cero). Definimos un ejemplo de transporte de media y varianza.

13. Evaluo el costo en este caso y realizo una aproximación de monte carlo utilizando las observaciones, obteniendo un funcional explícito a optimizar, donde podemos notar que el costo penaliza medias diferentes a 0 y varianzas de 1. Descompogo un kernel de covarianza en su función de varianza y su kernel de correlación, y menciono que media y varianza definen la marginal, la correlación define la copula.

14. Con esta descomposición, se obtiene que el transporte del kernel se puede expresar como la composición del transporte de varianza y el transporte de correlación (y conmutan). Defino un transporte aditivo de la media, y mostramos que la composición de estos tres transportes corresponde al transporte asociado a un GP desde un proceso de ruido blanco. Si consideramos el transporte lineal estocástico con fuente gamma inversa, entonces mostramos la composición de transportes que generan un Student-t process.

15. Muestro las descomposiciones de los transportes asociados a warped GP y warped Student-t, ambas agregando un el transporte inducido por un mappeo no lineal.

A continuación describiré cada uno de los puntos de trabajos futuros:

1. Agregar el caso de Skew Gaussian Processes.
2. Buscar otras transformadas que no sean triangulares, tal como es el caso GPMM.
3. Definir mejores costos según una perspectiva estadística o numérica, si hay un costo natural.
4. Familias de transportes que sea el problema ad-hoc, hasta donde se puede llegar (deep).
5. En termino de medida, el funcional a optimizar es convexo, el problema de la no convexidad aparece al parametrizar las distribuciones. Buscar si es posible aprovecharse de este hecho para encontrar mejores optimos £ puedo generar a partir de mis iteraciones una distribución mejor? Enfoque bayesiano y model average.
6. Revisar sobre flujo gradiente en el paper de Otto y Kinder del año 1998. Tal vez ayuda a encontrar un algoritmo de gradiente en el espacio de medida (W^2 u otro costo).
7. Ver la complejidad del transporte según la cantidad de fuentes aleatorias, como considerar dos gaussianas por cada coordenada.
8. Definir bien los conceptos de prior y posterior, dar noción de costo. Ver relación con BIC e AIK.
9. Revisar si el tema de kernel embedding cabe en este contexto.
10. Estudiar bien el paper de clasificación con transporte óptimo.
11. Revisar bien la razón de tomar transportes triangulares, si es por un tema algoritmico, numérico, geométrico, etc.