حل تمرین داده کاوی

(بخش دوم)

استاد:جناب آقای دکتر فراهانی

استاد یار: آقای شریفی

دانشجو: یا

حل تمرین داده کاوی

(بخش دوم)

استاد:جناب آقای دکتر فراهانی

استاد یار: آقای شریفی

دانشجو: یاشار موسی پور

شار موسی پور

8- Bootstrapping چیست و چه تفاوتی با Validation Cross دارد؟ در کجا ها از Bootstrappinاستفاده میشود ؟

پاسخ:

Bootstrapping یک روش باز نمونه گیری برای تخمین پارامترهای جامعه.

Validation Cross بررسی در توانایی الگوریتم برای فراگیری و تعمیم.

زمانی که بررسی بر مواردی از جامعه میباشد که اعضای جامعه به راحتی قابل شناسایی یا در دسترس نیستند.

به عنوان مثال

- بررسی در مورد یک گونه جانوری یا گیاهی کمیاب

- بیمارهای خاص که اعضای واقعی قابل شناسایی نیستند

10- در خصوص الگوریتم های مختلف ساخت درخت تصمیم (همانندID3 ، CARTو ...) تحقیق کنید و به صورت کلی مشخص نمایید تفاوت الگوریتم های مختلف ساخت درخت تصمیم در چیست ؟

پاسخ :

زمانی که درخت برای کارهای طبقه‌بندی استفاده می‌شود، به عنوان درخت طبقه‌بندی (Classification Tree)**شناخته می‌شود .**

و هنگامی که برای فعالیت‌های رگرسیونی به کار می‌رود درخت رگرسیون (RegressionDecisionTree) **نامیده می‌شود .**

انواع الگوریتم‌های شایع برای برقراری درخت تصمیم

- ID3

یکی از الگوریتم‌های بسیار ساده درخت تصمیم که در سال 1986 توسط Quinlan مطرح شده است. اطلاعات به دست آمده به عنوان معیار تفکیک به کار می‌رود. این الگوریتم هیچ فرایند هرس کردن را به کار نمی‌برد و مقادیر اسمی و مفقوده را مورد توجه قرار نمی‌دهد.

C4.5 -

این الگوریتم درخت تصمیم، تکامل یافته ID3 است که در سال 1993 توسط Quinlan مطرح شده است.

Gain Ratio –

به عنوان معیار تفکیک در نظر گرفته می‌شود. عمل تفکیک زمانی که تمامی نمونه‌ها پایین آستانه مشخصی واقع می‌شوند، متوقف می‌شود. پس از فاز رشد درخت عمل هرس کردن بر اساس خطا اعمال می‌شود. این الگوریتم مشخصه‌های اسمی را نیز در نظر می‌گیرد.

CART -

برای برقراری درخت‌های رگرسیون و دسته‌بندی از این الگوریتم استفاده می‌شود. در سال 1984توسط Breiman و همکارانش ارائه شده است. نکته حائز اهمیت این است که این الگوریتم درخت‌های باینری ایجاد می‌کند به طوری که از هر گره داخلی دو لبه از آن خارج می‌شود و درخت‌های بدست آمده توسط روش اثربخشی هزینه، هرس می‌شوند.

یکی از ویژگی‌های این الگوریتم، توانایی در تولید درخت‌های رگرسیون است. در این نوع از درخت‌ها برگ‌ها به جای کلاس مقدار واقعی را پیش‌بینی می‌کنند. الگوریتم برای تفکیک کننده‌ها، میزان مینیمم مربع خطا را جستجو می‌کند. در هر برگ، مقدار پیش‌بینی بر اساس میانگین خطای گره‌ها می‌باشد.

CHID -

این الگوریتم درخت تصمیم به جهت در نظرگرفتن مشخصه‌های اسمی در سال 1981 توسط Kass طراحی شده است. الگوریتم برای هر مشخصه ورودی یک جفت مقدار که حداقل تفاوت را با مشخصه هدف داشته باشد، پیدا می‌کند.

.....

۱۳ . 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؟

پاسخ: هرس کردن (Pruning) یک الگوریتم ساده و شهودی است. انواع مختلفی وجود دارد، اما ایده اصلی بر روی هر شبکه عصبی کار می‌کند. ایده این است. درون یک شبکه عصبی بزرگ آموزش‌دیده، مقداری وزن با اندازه بزرگ و مقداری با اندازه کوچک وجود خواهد داشت. به طور طبیعی، وزن‌های با اندازه بزرگ بیشتر به خروجی شبکه کمک می‌کنند. بنابراین، برای کاهش اندازه شبکه، ما از وزن‌های کوچک (هرس) خلاص می‌شویم.

یک مزیت هرس این است که واقعا خوب عمل می‌کند. شواهد تجربی قابل‌توجهی برای حمایت از آن وجود دارد. به طور خاص‌تر، هرس برای حفظ دقت و در عین حال کاهش اندازه شبکه (حافظه) نشان‌داده شده‌است.
