بسمه تعالی

درس داده‌کاوی

جواب تمرین های سری دوم

شایان ذکر است این فایل مربوط به جواب تمرین های تشریحی میباشد. تمرین هایی که نیاز به کدنویسی دارند به پیوست تقدیم میگردد.

نگارش:سپهر پیریائی (۹۹۴۲۲۰۴۰)

جناب آقای دکتر هادی فراهانی، جناب آقای دکتر خردپیشه

خرداد ۱۴۰۰

جواب مربوط به تمرین شماره۳:

قضیه بیز روشی برای دسته‌بندی پدیده‌ها، بر پایه احتمال وقوع یا عدم وقوع یک پدیده‌است و در نظریه احتمالات با اهمیت و پرکاربرد است. اگر برای فضای نمونه‌ای مفروضی بتوانیم چنان افرازی انتخاب کنیم که با دانستن اینکه کدامیک از پیشامدهای افراز شده رخ داده‌است، بخش مهمی از عدم قطعیت تقلیل می‌یابد. این قضیه از آن جهت مفید است که می‌توان از طریق آن، احتمال یک پیشامد را با مشروط کردن نسبت به وقوع یا عدم وقوع یک پیشامد دیگر محاسبه کرد. در بسیاری از حالت‌ها، محاسبه احتمال یک پیشامد به صورت مستقیم کاری دشوار است. با استفاده از این قضیه و مشروط کردن پیشامد مورد نظر نسبت به پیشامد دیگر، می‌توان احتمال مورد نظر را محاسبه کرد.

Gaussian Naive Bayes:

اگر داده‌ها از نوع پیوسته باشند، از مدل احتمالی با توزیع گاوسی یا نرمال برای متغیرهای مربوط به شواهد می‌توان استفاده کرد. در این حالت هر دسته یا گروه دارای توزیع گاوسی است. به این ترتیب اگر k دسته یا کلاس داشته باشیم می‌توانیم برای هر دسته میانگین و واریانس را محاسبه کرده و پارامترهای توزیع نرمال را برای آن‌ها برآورد کنیم. فرض کنید که μk میانگین و $\sigma 2k$ واریانس دسته k ام یعنی Ck باشد. همچنین v را مشاهدات حاصل از متغیرهای تصادفی X در نظر گرفت. از آنجایی که توزیع X در هر دسته گاوسی (نرمال) فرض شده است، خواهیم داشت:

$$p(x = v \mid C_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{(v-\mu_k)^2}{2\sigma_k^2}}$$

:

Multinomial Naive Bayes

بیز ساده چندجمله‌ای، به عنوان یک دسته‌بند متنی بسیار به کار می‌آید. در این حالت برحسب مدل احتمالی یا توزیع چند جمله‌ای، برداری از n ویژگی برای یک مشاهده به صورت $X=(x1,\ldots,xn)$ با احتمالات $(p1,\ldots,pn)$ در نظر گرفته می‌شود. مشخص است که در این حالت بردار X بیانگر تعداد مشاهداتی است که ویژگی خاصی را دارا هستند. به این ترتیب تابع درستنمایی در چنین مدلی به شکل زیر نوشته می‌شود:

$$p(\mathbf{x} \mid C_k) = \frac{(\sum_i x_i)!}{\prod_i x_i!} \prod_i {p_{ki}}^{x_i}$$

Bernoulli Naive Bayes:

در این قسمت به بررسی توزیع برنولی و دسته‌بندی بیز خواهیم پرداخت. به شکلی این نوع از دسته‌بند بیز بیشترین کاربرد را در دسته‌بندی متن‌های کوتاه داشته، به همین دلیل محبوبیت بیشتری نیز دارد. در این مدل در حالت چند متغیره، فرض بر این است که وجود یا ناموجود بودن یک ویژگی در نظر گرفته شود. برای مثال با توجه به یک لغتنامه مربوط به اصطلاحات ورزشی، متن دلخواهی مورد تجزیه و تحلیل قرار می‌گیرد و بررسی می‌شود که آیا کلمات مربوط به لغتنامه ورزشی در متن وجود دارند یا خیر. به این ترتیب مدل تابع درستنمایی متن براساس کلاس های مختلف Ck به شکل زیر نوشته می‌شود:

$$p(\mathbf{x} \mid C_k) = \prod_{i=1}^{n} p_{ki}^{x_i} (1 - p_{ki})^{(1-x_i)}$$

جواب مربوط به تمرین شماره ۱۳:

تعداد همسایه ها باعث تغییر در دقت یادگیری می‌شود. در الگوریتم k نزدیکترین همسایگی، classification به میزان بیشترین تعداد مشترک دسته بندی شده همسایگان می‌باشد. برای مثال اگر تعداد نزدیکترین همسایگی را همانند شکل زیر یکبار ۳ در نظر بگیریم، نتیجه با توجه به دو همسایه نزدیک قرمز و یک همسایه آبی، قرمز خواهد بود. ولی در صورتی که نزدیکترین همسایگی را ۵ در نظر بگیریم، نتیجه از نوع کلاس آبی خواهد بود. بنابرین تعداد میزان همسایگی در هر مساله متقاوت هست و معمولا بیشتر با آزمون و خطا می‌توان به نتایج دقیق‌تری دست یافت. البته باید به این نکته توجه کرد که هر چه تعداد همسایگان را بیشتر در نظر بگیریم، احتمال پراکندگی نتایج ممکن است بیشتر باشد. و اگر تعداد همسایگان نیز خیلی کم باشد، باعث خطا (با توجه به داده‌های استثنا) و خطی بودن نتایج شود.

جواب مربوط به تمرین شماره ۱۵:

در یک مدل پارامتری ، تعداد پارامترها با توجه به اندازه نمونه ثابت می شود. در یک مدل غیر پارامتری ، تعداد (موثر) پارامترها می توانند با اندازه نمونه رشد کنند.
در یک رگرسیون OLS ، تعداد پارامترها همیشه به طول β خواهد بود، به علاوه یک واریانس.
یک شبکه عصبی با معماری ثابت و بدون پوسیدگی وزن یک مدل پارامتریک است.
اما اگر دچار فروپاشی وزن هستید ، مقدار پارامتر پوسیدگی که با اعتبار سنجی متقابل انتخاب می شود ، با داده های بیشتر ، به طور کلی کوچکتر می شود. این می تواند به عنوان افزایش تعداد موثر پارامترها با افزایش اندازه نمونه تفسیر شود.

جواب مربوط به تمرین شماره ۱۶:

Matthews Correlation Coefficient:

پارامتری است که برای ارزیابی کارایی الگوریتم‌های یادگیری ماشین از آن استفاده می‌شود. این پارامتر بیان‌گر کیفیت کلاس‌بندی برای یک مجموعه باینری می‌باشد. بنابرین مواقعی از این معیار استفاده می‌گردد که classification ما همیشه دو بخشی باشد.