به نام خدا

دانشگاه تهران

پردیس دانشکده‌های فنی

دانشکده برق و کامپیوتر

سیستم‌های هوشمند

تمرین شماره ۱

فرزاد مهری

۸۱۰۱۹۴۴۱۰

سوال ۱

الف)

بدون استفاده از ویژگی های داده شده، احتمال بیمار بودن را میتوان با نسبت افراد بیمار به کل افراد تقریب زد. با این روش احتمال بیمار بودن به صورت روبرو بدست آمد:

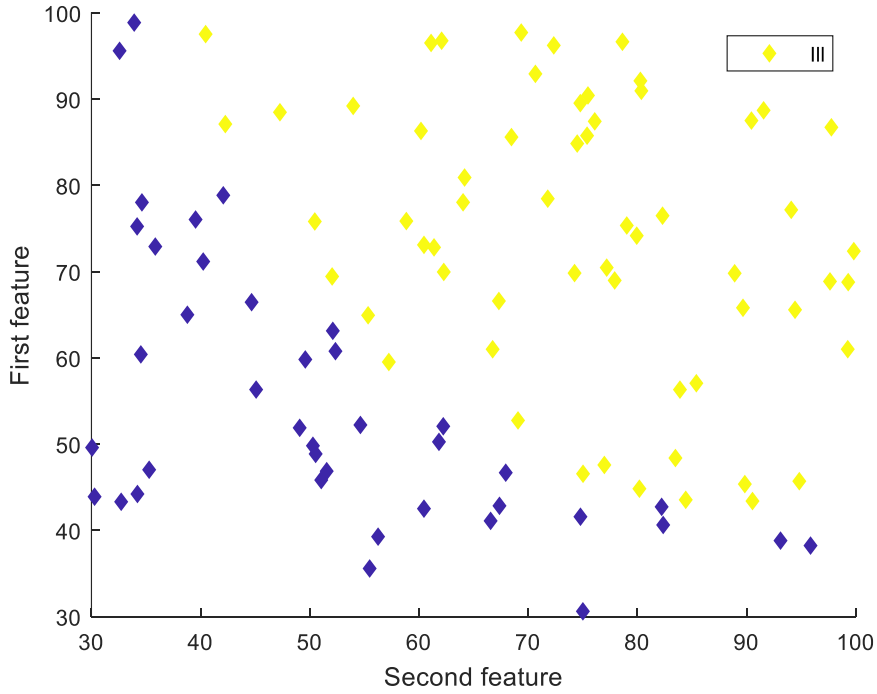
$$P_{illness} = 0.2250$$

ب)

با در نظر گرفتن توضیع نرمال برای پارامتر ها و تخمین آن ها با روش Naïve Bayes، بیمار بودن داده های تست با دقت ۰,۹۶۳۰ تخمین زده شد.

سوال ۲

الف)



شکل ۱، داده های ابی رنگ مربوط به افراد سالم و داده های زرد مربوط به افراد بیمار است.

ب)

رابطه ی گرادیان تابع هزینه:

$$\frac{\partial}{\partial\theta}J = \frac{1}{m}\sum\frac{\partial}{\partial\theta}\left(-y^{(i)}\log(h_\theta(x^{(i)}) - (1-y^{(i)})(1-\log(h_\theta(x^{(i)})\right)$$
$$= -\frac{1}{m}\sum\left[y^{(i)}\frac{\partial}{\partial\theta}\left(\log(h_\theta(x^{(i)})\right) - (1-y^{(i)})\frac{\partial}{\partial\theta}\left(\log(1-h_\theta(x^{(i)})\right)\right]$$
$$= -\frac{1}{m}\sum\left[y^{(i)}\frac{\partial}{\partial\theta}\left(\log\left(\frac{1}{1+e^{-\theta^T x^{(i)}}}\right)\right)\right.$$
$$\left.+ (1-y^{(i)})\frac{\partial}{\partial\theta}\left(\log(1-\frac{1}{1+e^{-\theta^T x^{(i)}}}\right)\right]$$
$$= -\frac{1}{m}\sum\left[y^{(i)}\left(-\frac{x^{(i)T}e^{-\theta^T x}}{1+e^{-\theta^T x^{(i)}}}\right)\right.$$
$$\left.+ (1-y^{(i)})\left(-\frac{x^{(i)T}}{e^{-\theta^T x^{(i)}}} + \frac{x^{(i)T}}{1+e^{-\theta^T x^{(i)}}}\right)\right]$$

پارامتر های θ با استفاده از روش گرادیان نزولی، در نهایت به صورت زیر بدست آمد (داده ها به ۵ قسمت تقسیم شدند و در هر مرحله ۱/۵ داده ها به عنوان تست در نظر گرفته شد و در هر کدام از این حالت ها پارامتر ها و خطا محسابه شد و در نهایت میانگین آن ها به عنوان مقدار نهایی انتخاب گردید):

$$\theta = \begin{matrix}-4.8787\\0.0457\\0.0388\end{matrix}$$ و دقت مدل برابر ۸۸٪ بدست آمد. مقدار -۴,۸۷۸۷ ضریب پارامتر بایاس است.

۳)

با افزودن نرم ۲ به تابع هزینه، پارامتر ها به صورت زیر محاسبه شد:

$$\theta = \begin{matrix}-0.0704\\-0.0123\\0.0003\end{matrix}$$ و دقت مدل برابر ۷۴٪

با افزودن نرم ۲، دقت کاهش یافت اما با محدود کردن فضای پارامتر ها، از over-fit جلوگیری میشود و کمک میکند پارامتر ها در حالت کلی تر صحیح باشند.

سوال ۳

الف)

با استفاده از معیار فاصله Euclidean، خطا برای k های مختلف به صورت زیر بدست آمد:

$$k = 3: 5.77\%, k = 5: 3.85\%, k = 7: 4.99\%, k = 9: 4.99\%$$

کمترین میزان خطا به ازای k=5 بدست آمد.

ب)

با انتخاب k=5 و معیار های فاصله Manhattan و Minkowski خطا به صورت زیر بدست آمد:

$$Manhattan: 5.77\%$$

$$Minkowski(q = 5): 7.05\%$$

ج)

یکی از روش های بهبود عملکرد KNN، اعمال نوعی نرمالیزاسیون بر روی داده هاست به صورتی که feature های مختلف، رنج یکسانی داشته باشند و قابل مقایسه شوند.

سوال ۴

$$L(\mu: X_i) = p_r(X|\mu) = \prod p_r(X_i|\mu) = \prod \frac{1}{X_i \sigma \sqrt{(2\pi)}} e^{-\frac{(lnX_i - \mu)^2}{2\sigma^2}}$$

$$\Rightarrow loglikelihood = \log(L) = \sum \log\left(\frac{1}{\sigma\sqrt{(2\pi)}}\right) - \log(X_i) - \frac{(lnX_i - \mu)^2}{2\sigma^2}$$

$$\frac{\partial}{\partial \mu} loglikelihood = 0 \Rightarrow \sum lnX_i - \mu = 0 \Rightarrow \mu = \frac{\sum_{i=1}^{n} lnX_i}{n}$$