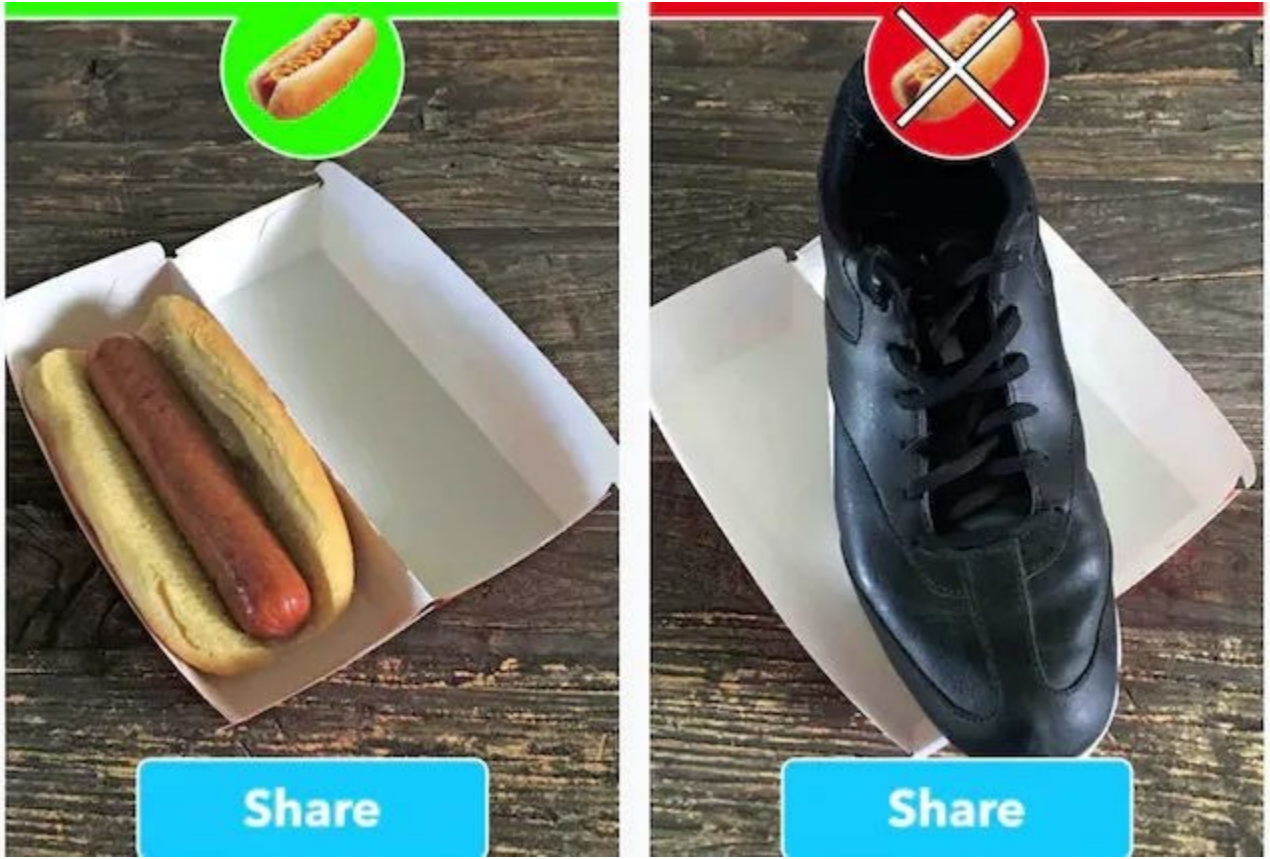
Классификацияға арналған сызықтық модельдер

Классификация



https://www.theverge.com/2017/6/26/15876006/hot-dog-app-android-silicon-valley

Логистикалық регрессия

ML рецепті

1. Кейбір үлгіні ойлап көріңіз
2. Деректерді үлгіге енгізіп, болжам жасаңыз.
3. Болжамдар мен шынайы мәндер арасындағы шығынды есептеңіз.
4. Ең аз шығын келтіретін үлгі параметрлерін анықтаңыз.

Логистикалық регрессия – үлгі

- Регрессияға арналған сызықтық моделімізді еске түсіріңіз

$$\hat{y}_i = \sum_j^p x_{ij} \cdot \beta_j$$

- Біз мұны жіктеу үшін пайдалана алмаймыз, өйткені біз 0 немесе 1 болжауымыз керек.
- Олай болса, мұны 0 мен 1 арасында сквош етейік.

Логистикалық регрессия - үлгі

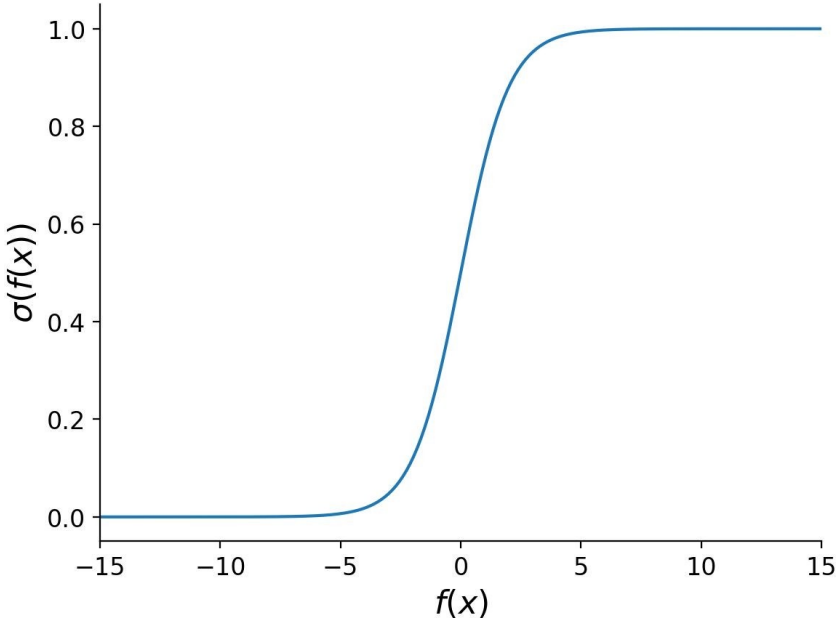
- Сызықтық үлгіні 0 мен 1 арасында сквош.

$$\hat{y}_i = \sum_j^p x_{ij} \cdot \beta_j$$

$$f(x_i) = \sum_j^p x_{ij} \cdot \beta_j$$

$$\hat{y}_i = \sigma(f(x_i))$$

$$\sigma(f(x_i)) = \frac{1}{1 + e^{-f(x_i)}}$$

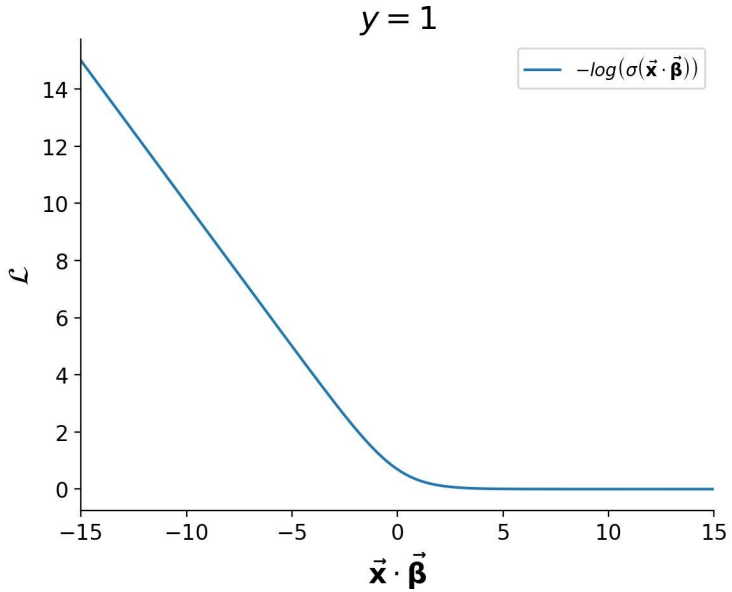


ML рецепті

1. Кейбір үлгіні ойлап көріңіз
2. Деректерді үлгіге енгізіп, болжам жасаңыз.
3. Болжамдар мен шынайы мәндер арасындағы шығынды есептеңіз.
4. Ең аз шығын келтіретін үлгі параметрлерін анықтаңыз.

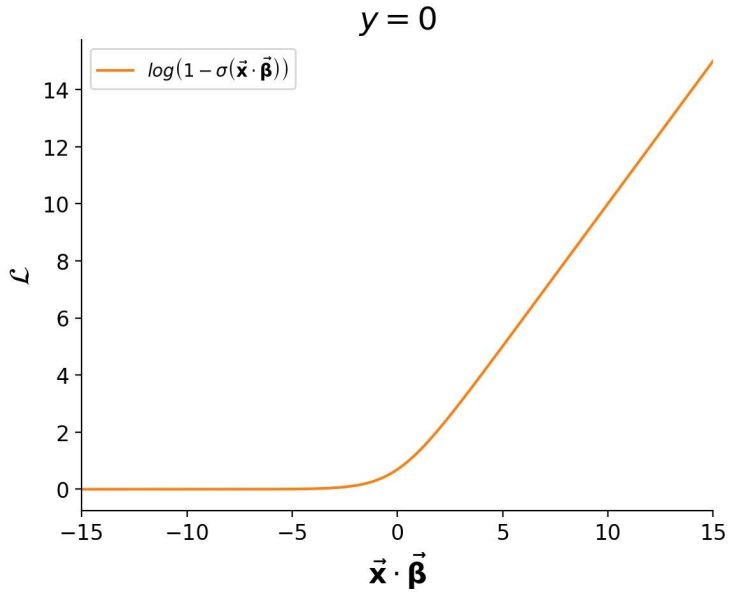
Логистикалық регрессия - жоғалту

$$\mathcal{L}\left(\vec{\beta}\right) = -\sum_{i=1}^{N} y_i \, log\left(\sigma\left(\vec{\mathbf{x}_\mathbf{i}} \cdot \vec{\beta}\right)\right)$$



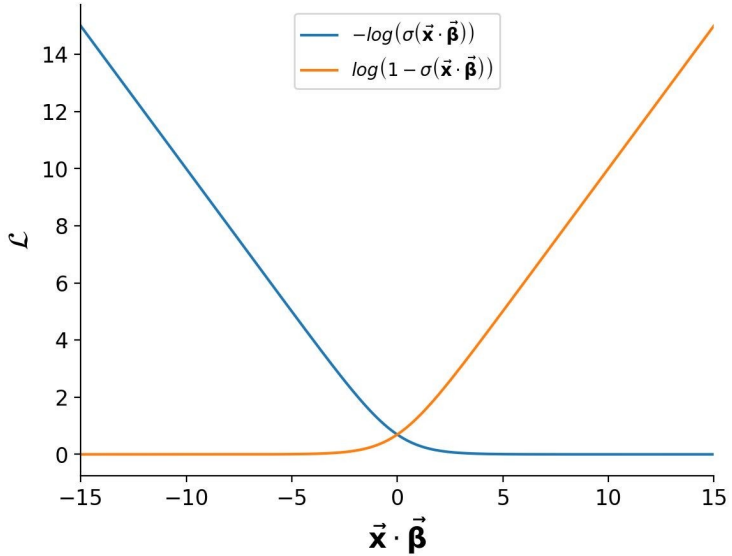
Логистикалық регрессия - жоғалту

$$\mathcal{L}\left(\vec{\beta}\right) = -\sum_{i=1}^{N} y_i \, log\left(\sigma\left(\overrightarrow{\mathbf{x_i}} \cdot \vec{\beta}\right)\right) + (1 - y_i) \, log\left(1 - \sigma\left(\overrightarrow{\mathbf{x_i}} \cdot \vec{\beta}\right)\right)$$



Логистикалық регрессия - жоғалту

$$\mathcal{L}\left(\vec{\beta}\right) = -\sum_{i=1}^{N} y_i \, log\left(\sigma\left(\vec{\mathbf{x}_\mathbf{i}} \cdot \vec{\beta}\right)\right) + (1 - y_i) \, log\left(1 - \sigma\left(\vec{\mathbf{x}_\mathbf{i}} \cdot \vec{\beta}\right)\right)$$



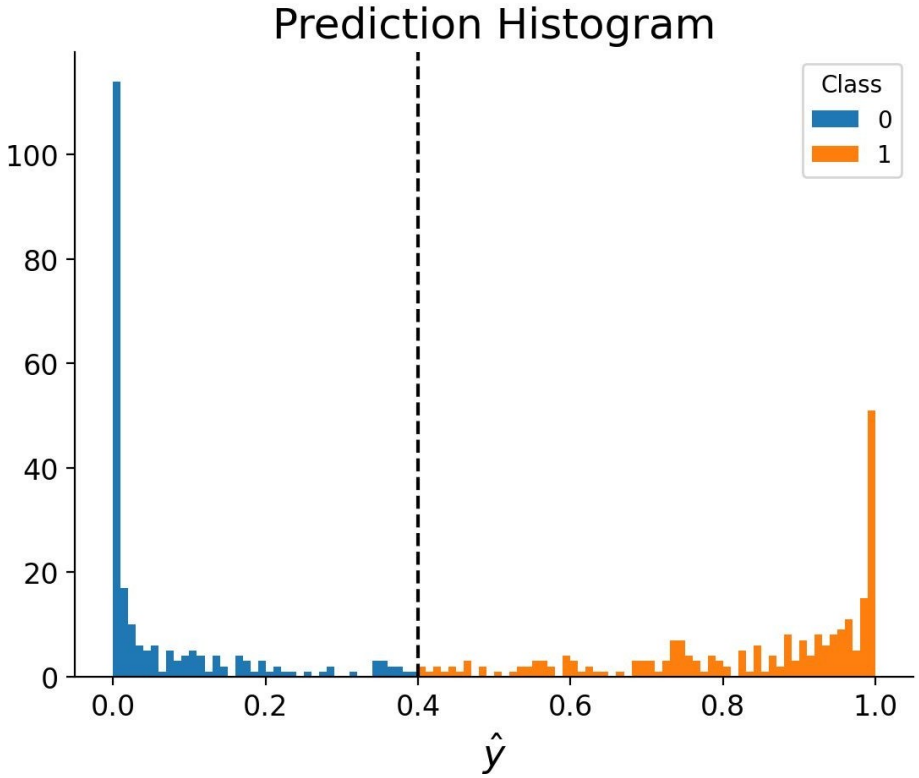
The ML Recipe

1. Кейбір үлгіні ойлап көріңіз
2. Деректерді үлгіге енгізіп, болжам жасаңыз.
3. Болжамдар мен шынайы мәндер арасындағы шығынды есептеңіз.
4. Ең аз шығын келтіретін үлгі параметрлерін анықтаңыз.

$$\frac{\partial \mathcal{L}}{\partial \vec{\beta}} = \sum_{i=1}^{N} \left(y_i - \frac{1}{1 + e^{\left(\vec{\mathbf{x}_i} \cdot \vec{\beta}\right)}} \right) \vec{\mathbf{x}_i}$$

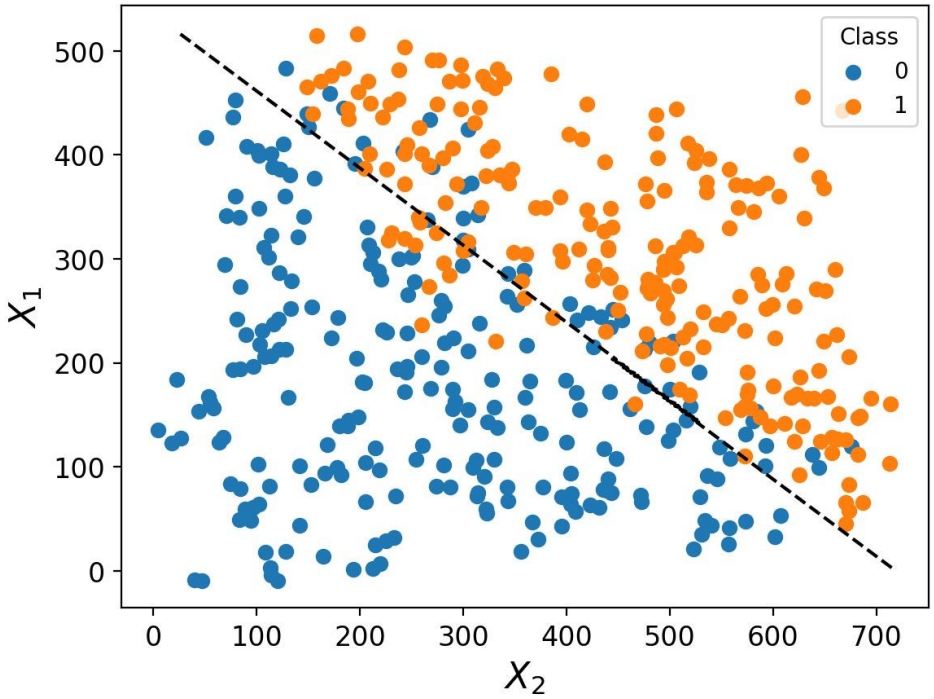
Логистикалық регрессия - Болжамдар

- Болжамдар 0 мен 1 аралығында.
- Оларды ықтималдық ретінде түсіндіруге болады.
- Ықтималдықтарды сыныптарға жіктеу үшін **шекті** таңдаймыз.



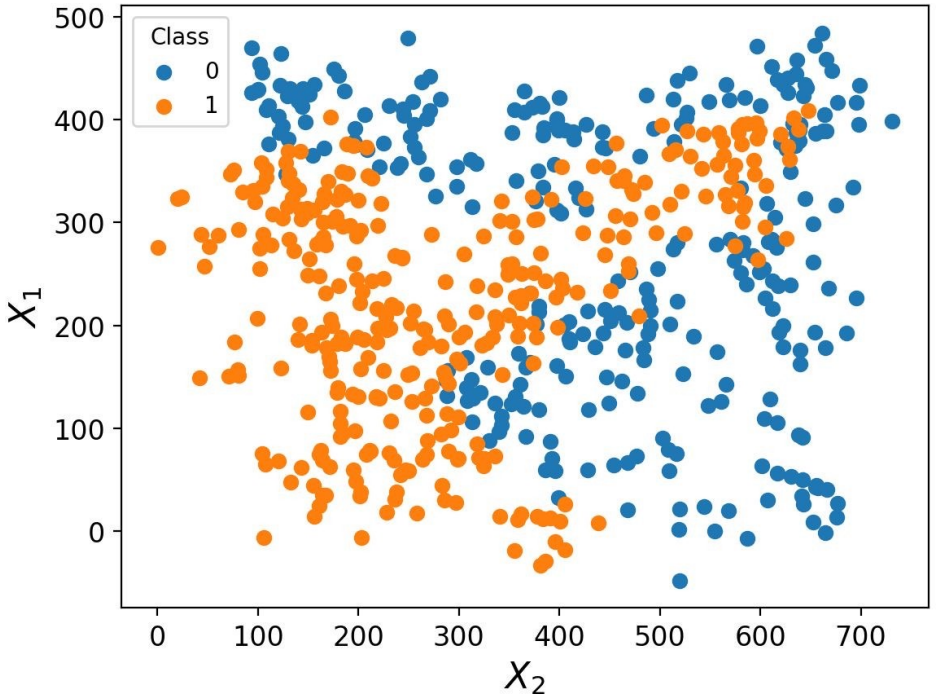
Логистикалық регрессия - шешімнің шекарасы

- 2 мүмкіндігі бар жіктеу мәселесін елестетіңіз.
- X1 және X2 кез келген мәндері үшін біздің
- модель 0 мен 1 арасындағы ықтималдықты болжайды.
- Шектен жоғары (төмен) болжамдар 1 (0) ретінде болжанады.
- Болжамдар аймағының бойымен өтетін сызық, онда болжау == шегі шешім шекарасы болып табылады.



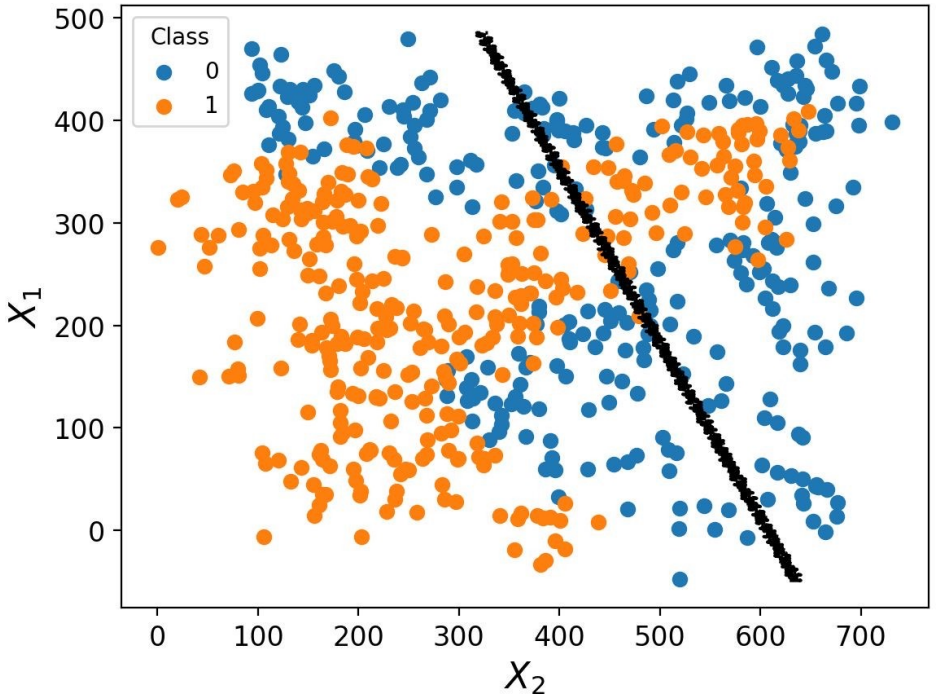
Логистикалық регрессия - шешімнің шекарасы

- Мәселелер көбінесе сызықты емес.



Логистикалық регрессия – шешімнің шекарасы

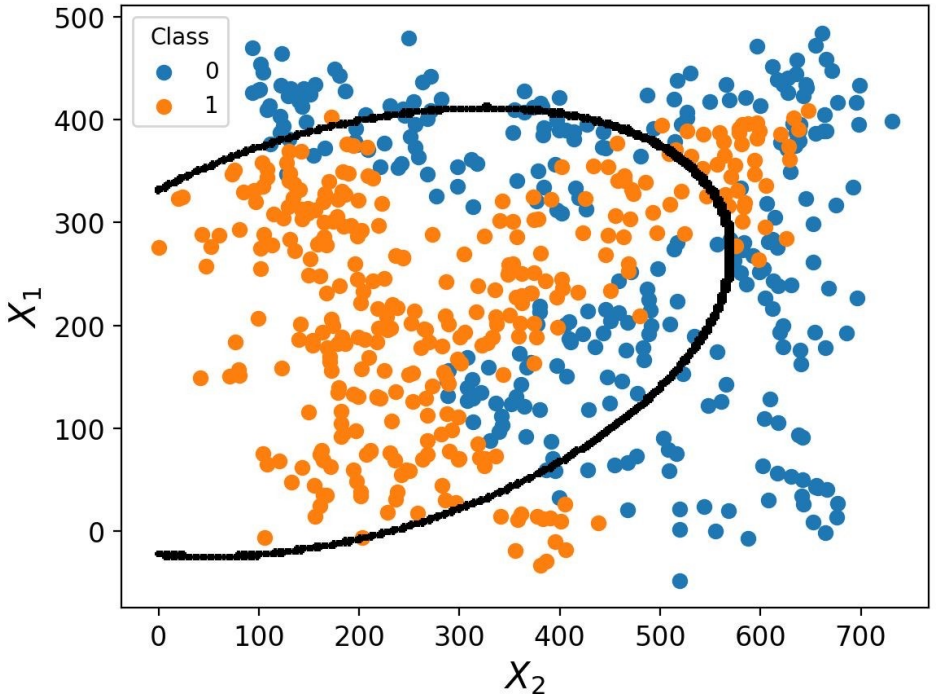
- Мәселелер көбінесе сызықты емес.
- Ал сызықтық шешім шекаралары оларда жақсы жұмыс істемейді!



Логистикалық регрессия – шешімнің шекарасы

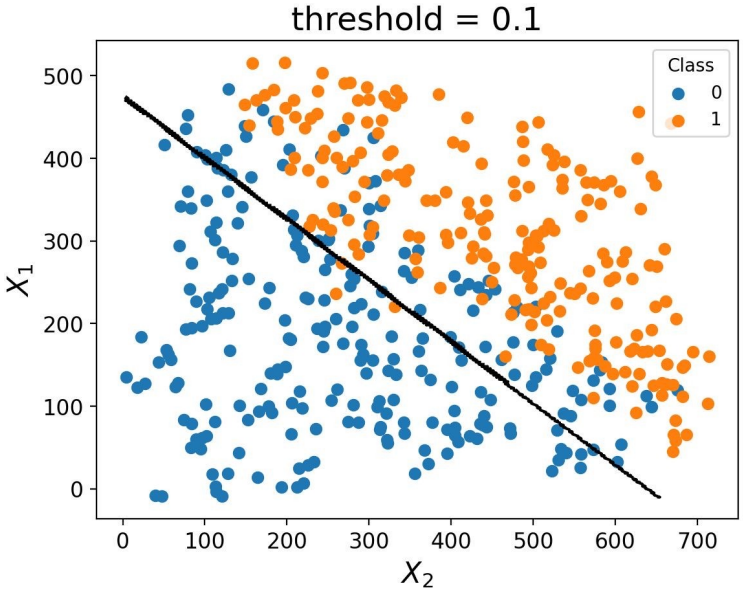
- Мәселелер көбінесе сызықты емес.
- Ал сызықтық шешім шекаралары оларда жақсы жұмыс істемейді!
- Сызықты емес мүмкіндіктерді жасауға болады:

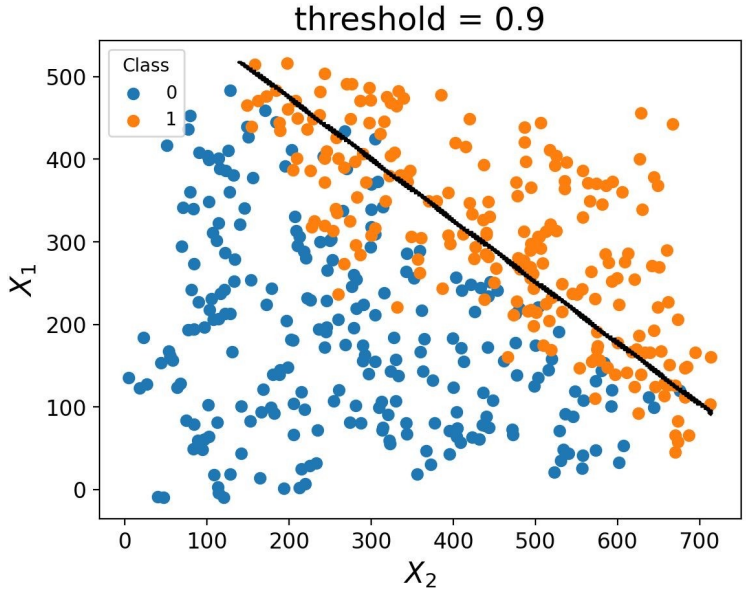
$$\hat{y}_i = \beta_0 + \beta_1 \cdot X_{i1} + \beta_2 \cdot X_{i2} + \beta_3 \cdot X_{i1} \cdot X_{i2} + \beta_4 \cdot X_{i1}^2 + \beta_5 \cdot X_{i2}^2$$



Логистикалық регрессия - шешімнің шекарасы

Сіздің табалдырық/шешім шекарасы – таңдау!



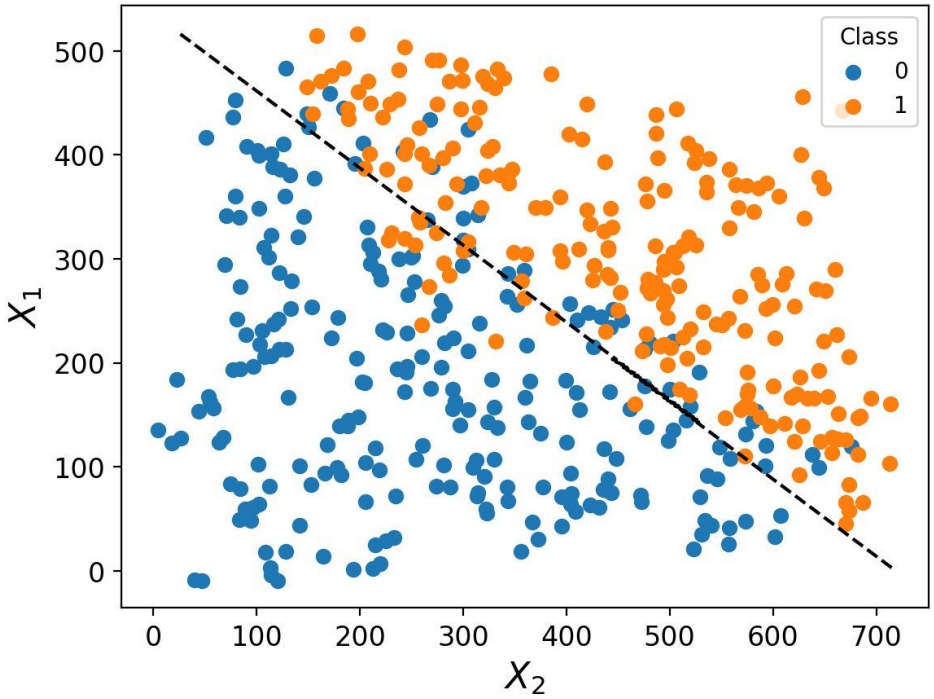


Классификациялық бағалау

Классификациялық бағалау

Шатасу матрицасы (confusion matrix)

Predicted Labels / True Labels	False	True
False	204	41
True	16	198

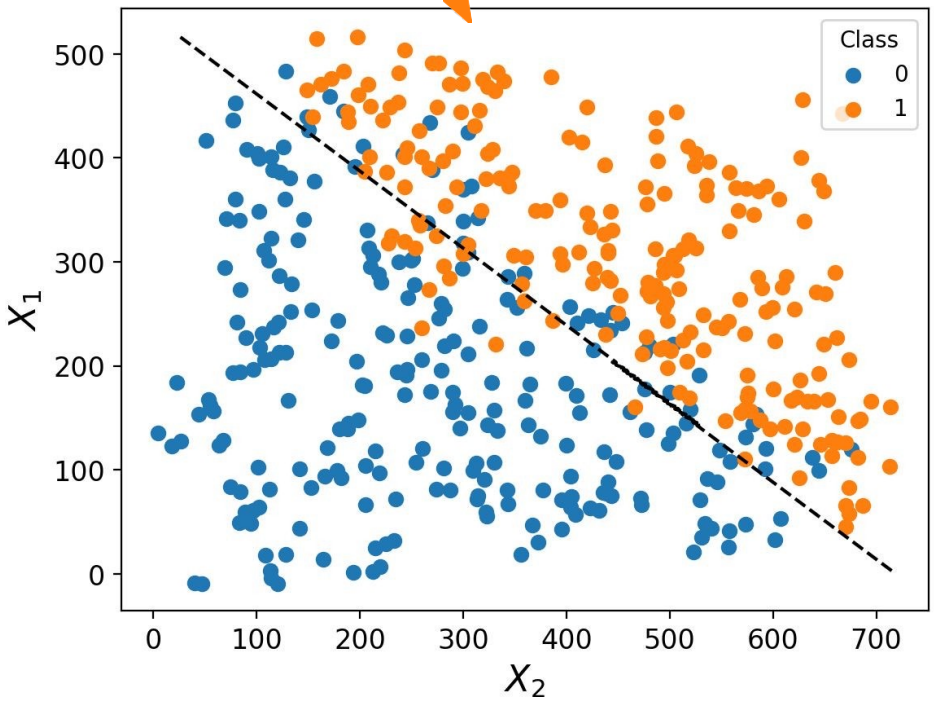


Классификациялық бағалау

Confusion Matrix

Predicted Labels / True Labels	False	True
False	204	41
True	16	198

True Positives
Осы жағындағы
барлық қызғылт
сары нүктелер

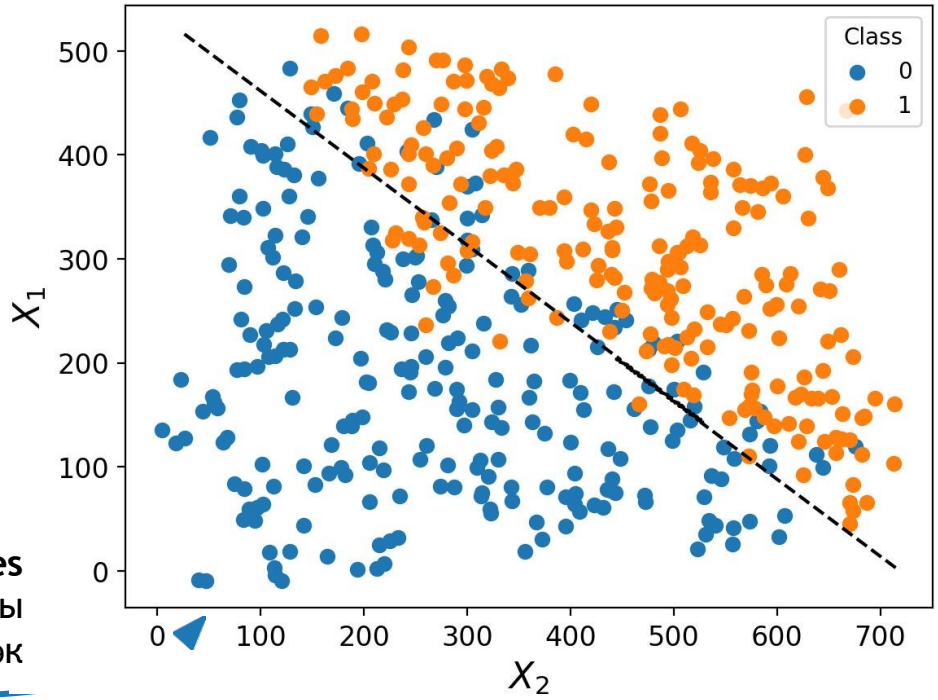


Классификациялық бағалау

Confusion Matrix

Predicted Labels / True Labels	False	True
False	204	41
True	16	198

True Negatives
Осы жағындағы барлық көк нүктелер



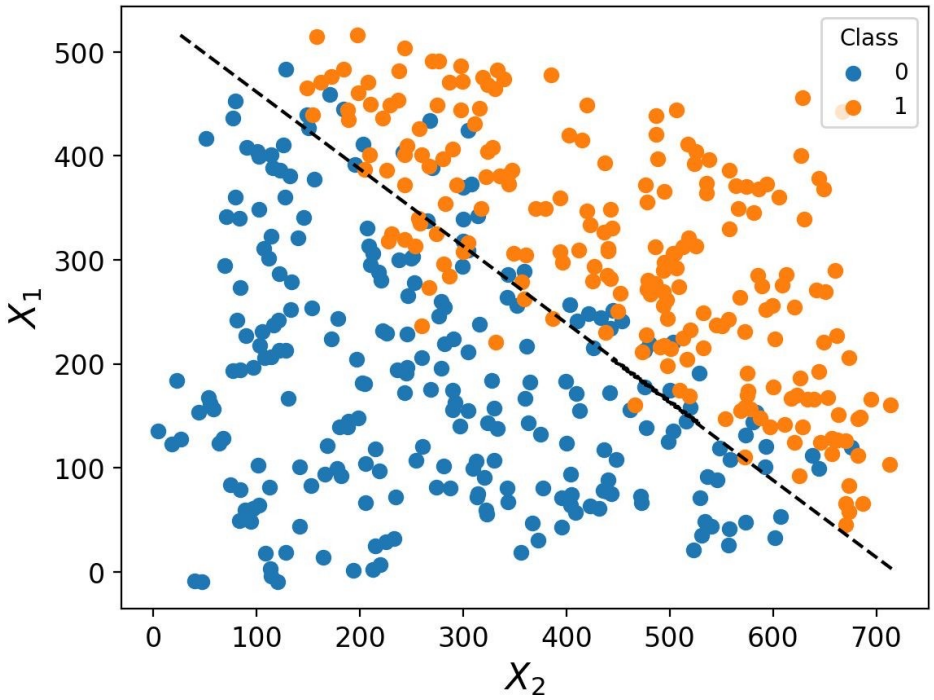
Классификациялық бағалау

Confusion Matrix

Predicted Labels / True Labels	False	True
False	204	41
True	16	198

False Positives

Осы жағындағы барлық көк нүктелер

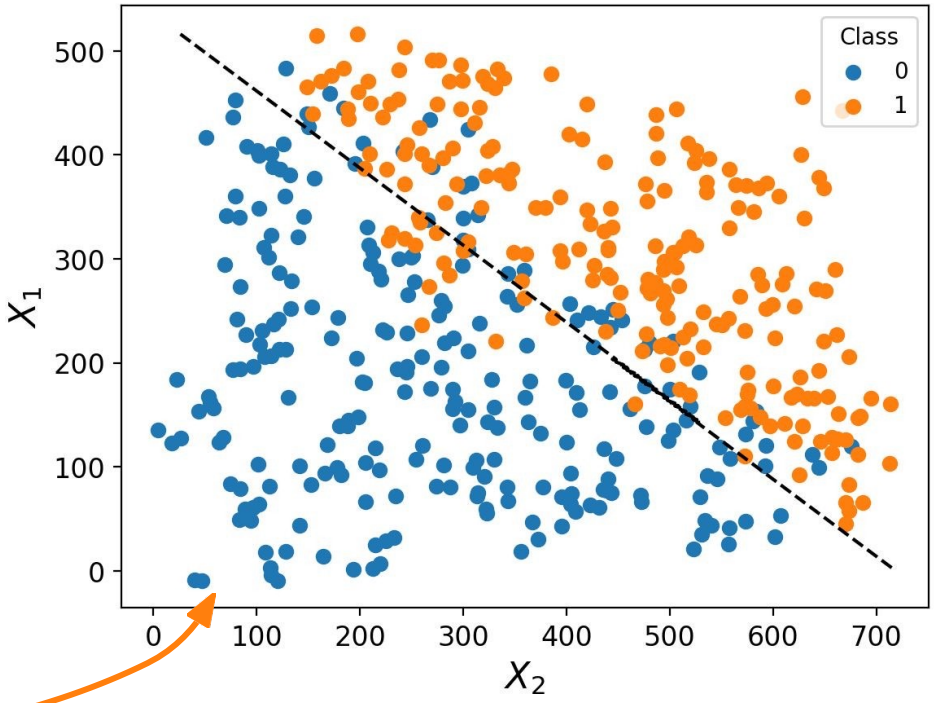


Классификациялық бағалау

Confusion Matrix

Predicted Labels / True Labels	False	True
False	204	41
True	16	198

False Negatives
Осы жағындағы барлық қызғылт сары нүктелер



Классификациялық бағалау

Confusion Matrix

Predicted Labels / True Labels	False	True
False	204	41
True	16	198

$$accuracy = \frac{TP + TN}{N}$$

Жалпы деретер

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

Классификациялық бағалау

Confusion Matrix

Predicted Labels / True Labels	False	True
False	204	41
True	16	198

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

Дәлдік: үлгілердің барлығы үшін менің моделім оң деп болжаған, олардың қанша пайызы шын мәнінде оң болды?

Еске салыңыз: Менің деректерімдегі барлық оң мысалдардың ішінде менің моделім неше % оң болжады?

Классификациялық бағалау

Confusion Matrix

Predicted Labels / True Labels	False	True
False	204	41
True	16	198

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

F1 ұпайы: дәлдік пен еске түсірудің гармоникалық ортасы

$$F_1 = \frac{2}{\frac{1}{precision} + \frac{1}{recall}}$$

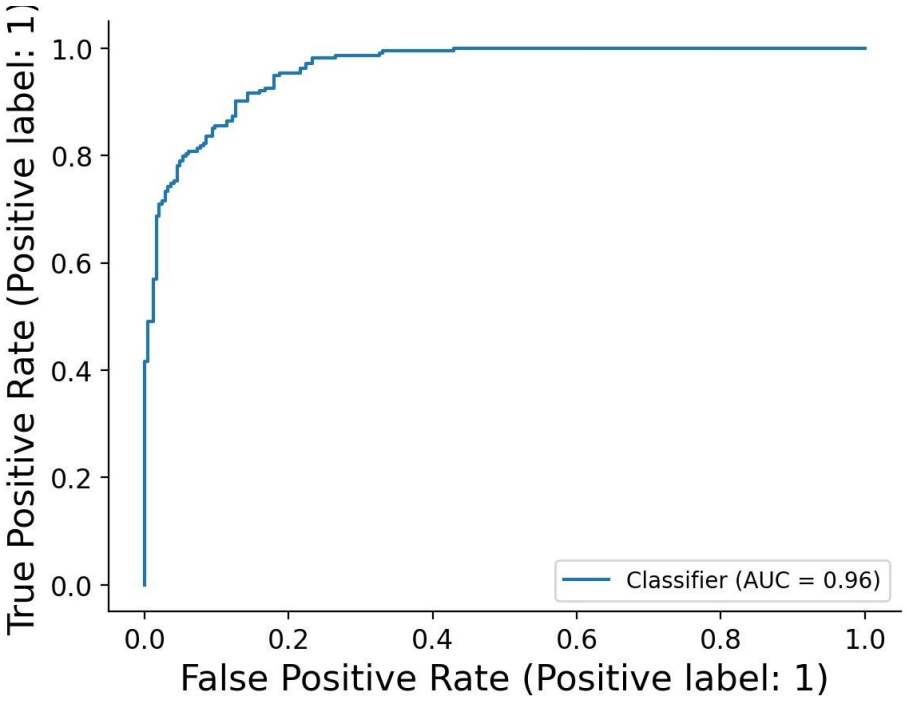
$$= 2\frac{precision \cdot recall}{precision + recall}$$

The ML Recipe

1. Кейбір үлгіні ойлап көріңіз
2. Деректерді үлгіге енгізіп, болжам жасаңыз.
3. Болжамдар мен шынайы мәндер арасындағы шығынды есептеңіз.
4. Ең аз шығын келтіретін үлгі параметрлерін анықтаңыз.
5. Жіктеу үлгісі болса, шекті таңдаңыз.
    1. Табалдырық – таңдау!
    2. Бұл таңдау сіздің үлгіңіздің бөлігі болып табылады.
    3. Шатасу матрицасы бір шекті мәнге негізделген.

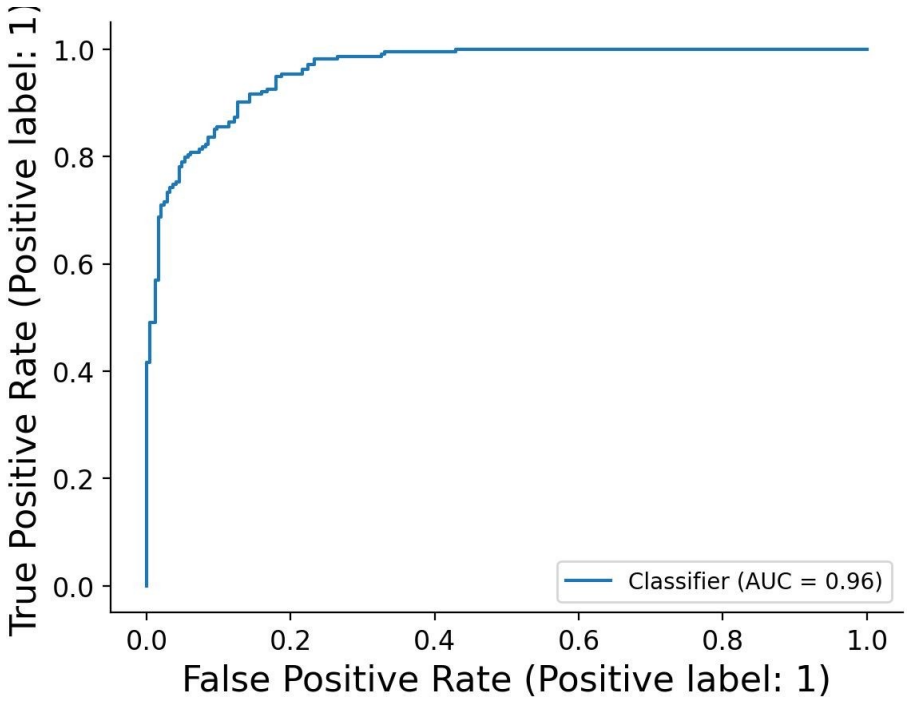
Классификациялық бағалау - шекті емес өлшемдер

- Қабылдағыштың жұмыс сипаттамасы (ROC) қисығы астындағы аудан
- (Кейде жай ғана AUC деп аталады)
- TPR = Recall = TP / (TP + FN)
- FPR = TP / (FP + TN)
- Жоғары және солға қарай жақсы
- Әрбір нүкте әртүрлі шекке сәйкес келеді
- Әдетте, шамамен ойыс.



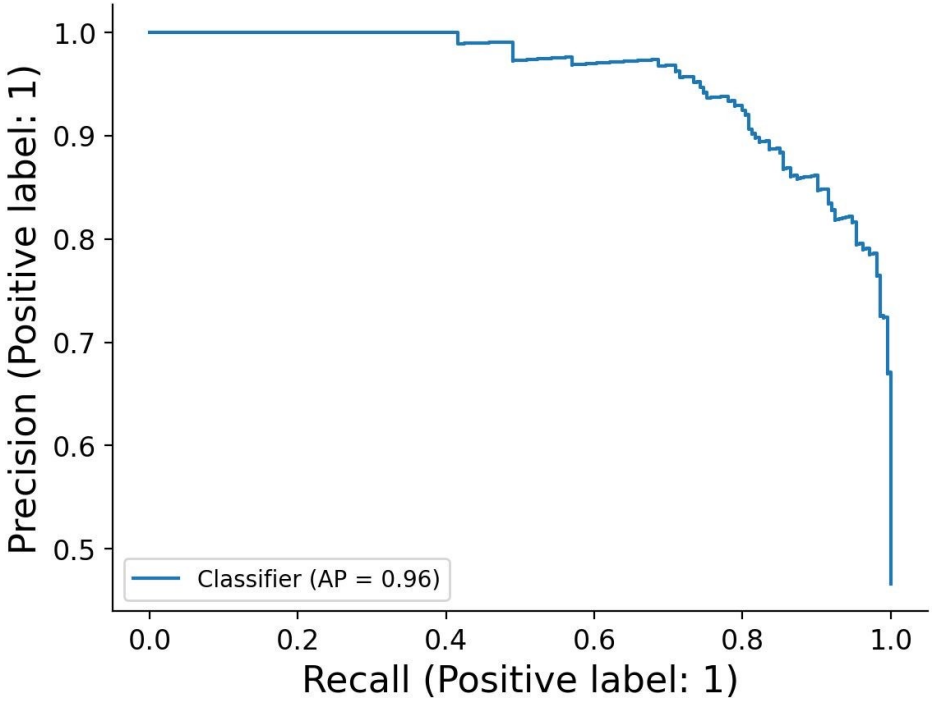
Классификациялық бағалау – шекті емес өлшемдер

- Негізгі AUROC – 0,5.
- яғни кездейсоқ классификатор
- Жақсы, үлгі өнімділігінің жалпы өлшемі.
- Теңгерімсіздік кластарының әсерінің көп бөлігін азайтады.
- «Егер сіз кездейсоқ 1 pos және 1 теріс үлгіні салсаңыз, оларды дұрыс дәрежелеу ықтималдығы қандай?» деп түсіндіре алады.



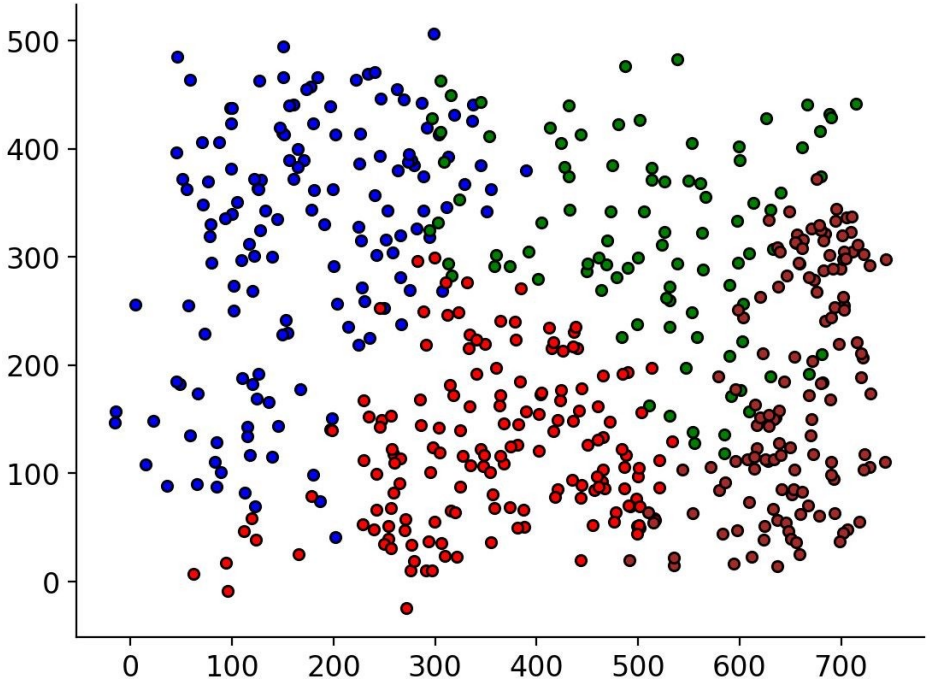
Классификациялық бағалау - шекті емес өлшемдер

- Дәлдік қайта шақыру қисығының астындағы аумақ (PR қисығы)
- «Орташа дәлдік»
- Жоғары және оңға қарай жақсы.
- Нашар модельдер өте ойыс емес PR қисықтарын көрсетеді
- Шекті таңдау үшін пайдалы.

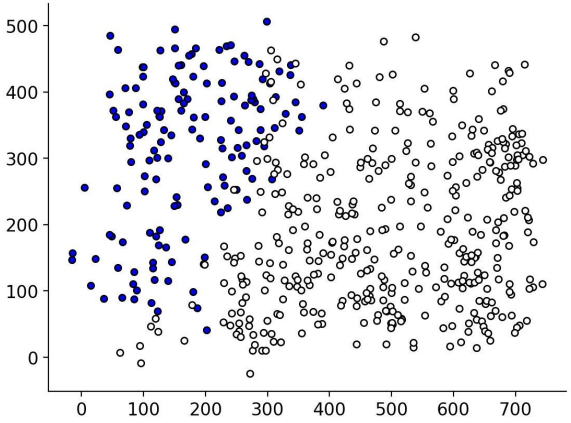


Көп класты классификация

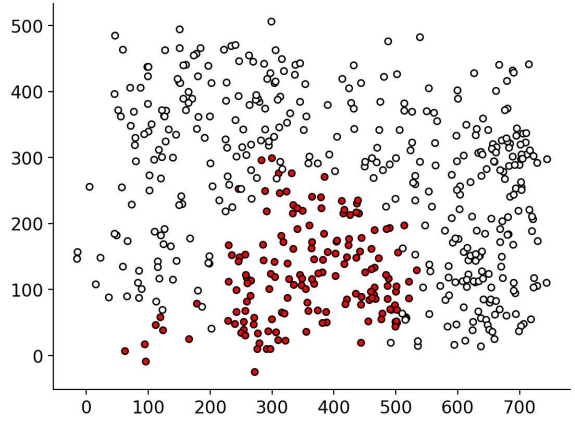
Көп класты классификация

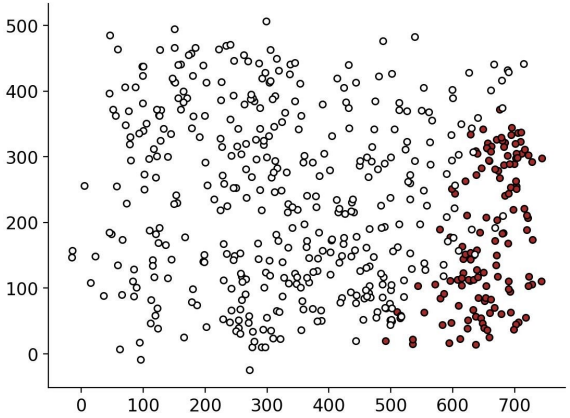


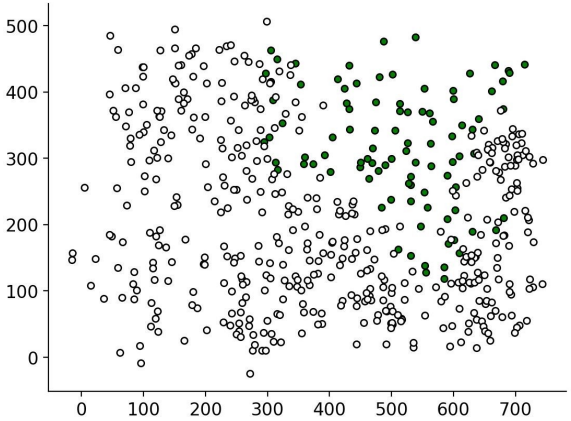
Көп класты жіктеу – one vs. all



- N сыныптары үшін N екілік классификаторларды жаттықтырыңыз.
- Әрбір N-сыныпты pos белгісі ретінде, қалған барлық сыныптарды теріс белгі ретінде қарастырыңыз.

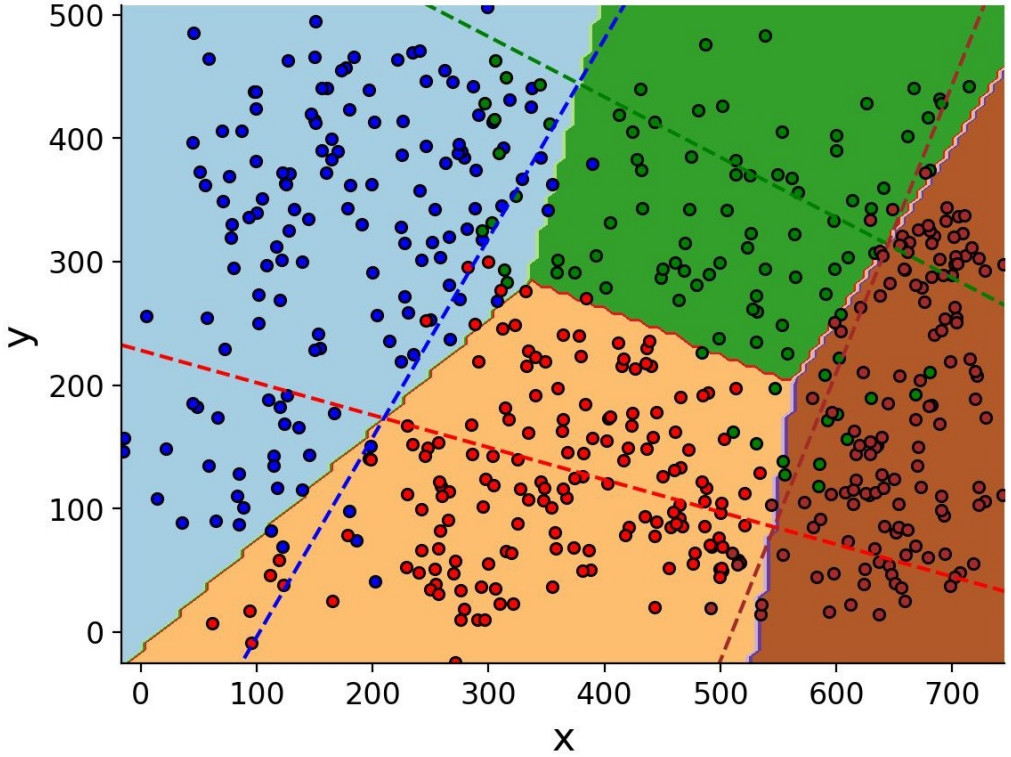






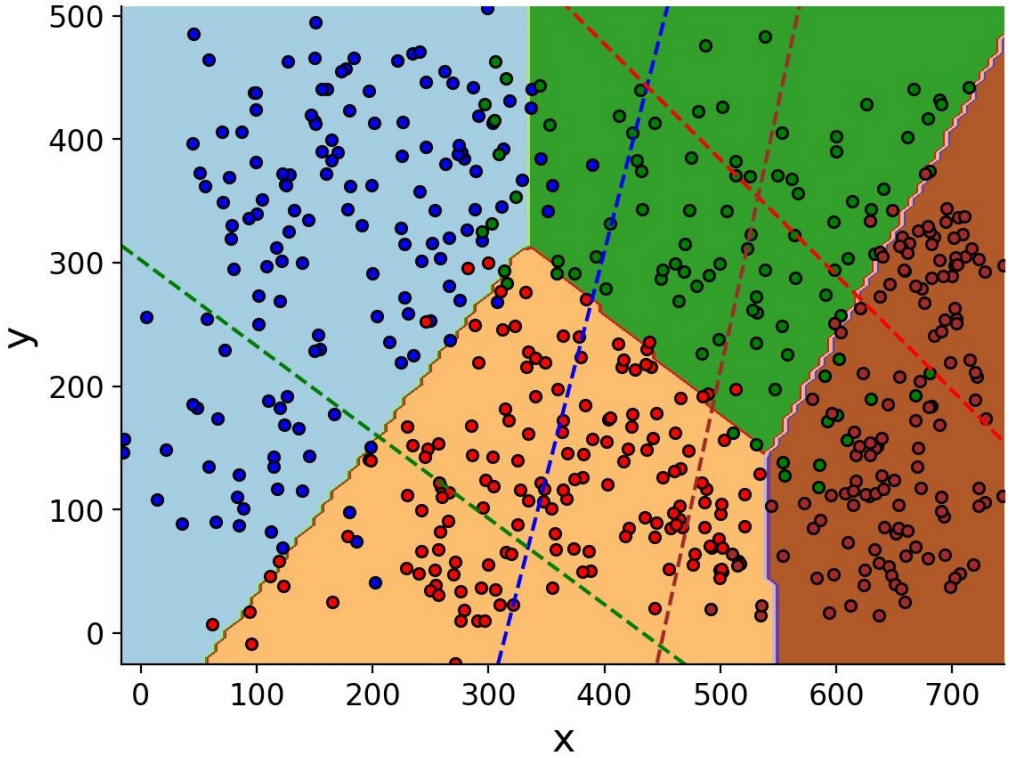
Көп класты жіктеу – one vs. rest

- Жіктеуіштің максималды баллы болжамды белгіні береді.



Көп кластық классификация - Көпмүшелік

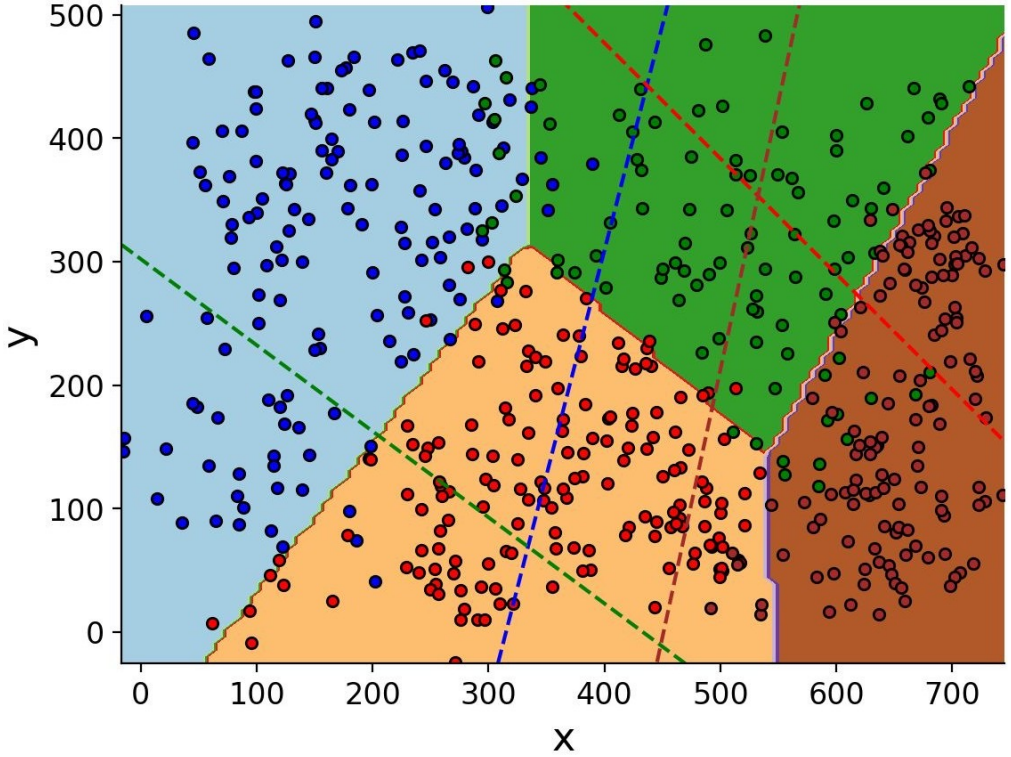
- Софтмакс немесе кросс энтропия деп те аталады



Көп кластық классификация - Көпмүшелік / Кросс энтропия

$$\mathcal{L}_{C\mathcal{E}} = \sum_{i=1}^{N} \left(- \sum_{c=1}^{C} y_{ic} log \left(s \left(\vec{\mathbf{x}_\mathbf{i}} \cdot \vec{\beta_\mathbf{c}} \right) \right) \right)$$

$$s \left(\vec{\mathbf{x}_\mathbf{i}} \cdot \vec{\beta_{\mathbf{c}'}} \right) = \frac{e^{\vec{\mathbf{x}_\mathbf{i}} \cdot \vec{\beta_{\mathbf{c}'}}}}{\sum_{c=1}^{C} e^{\vec{\mathbf{x}_\mathbf{i}} \cdot \vec{\beta_\mathbf{c}}}}$$



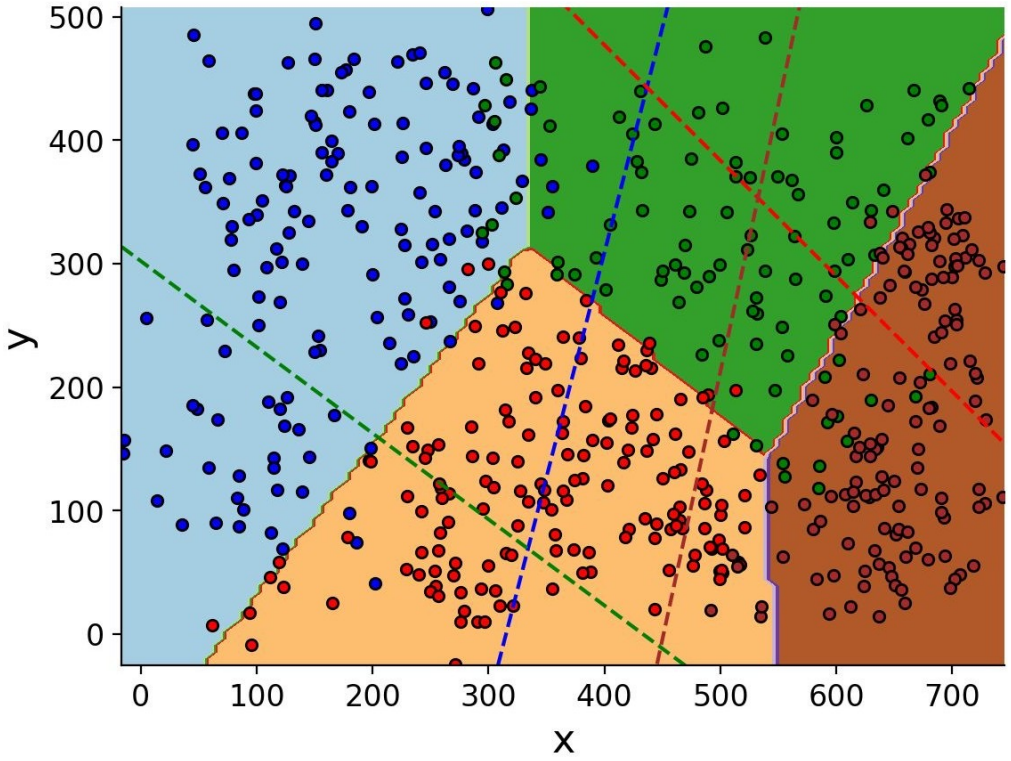
Көп кластық классификация - Көпмүшелік / Кросс энтропия

$$\mathcal{L}_{C\mathcal{E}} = \sum_{i=1}^{N} \left(- \sum_{c=1}^{C} y_{ic} log \left(s \left(\vec{\mathbf{x}_\mathbf{i}} \cdot \vec{\beta_\mathbf{c}} \right) \right) \right)$$

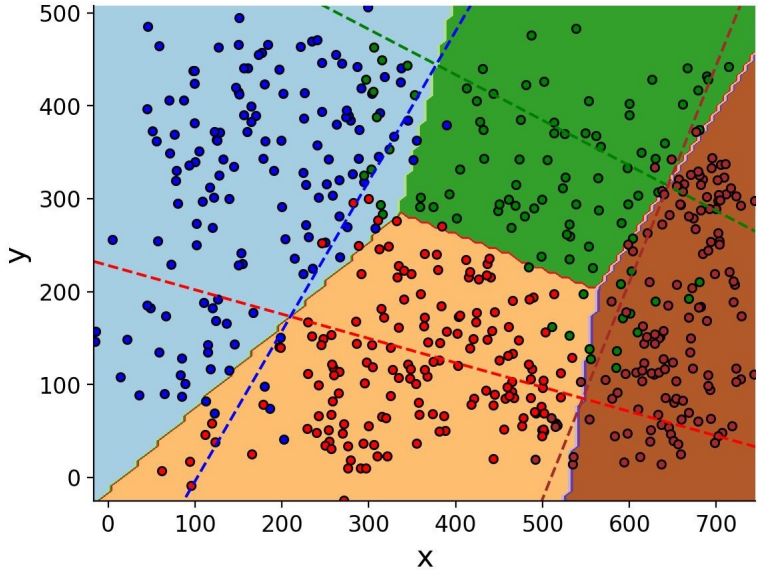
$$s \left(\vec{\mathbf{x}_\mathbf{i}} \cdot \vec{\beta_{\mathbf{c}'}} \right) = \frac{e^{\vec{\mathbf{x}_\mathbf{i}} \cdot \vec{\beta_{\mathbf{c}'}}}}{\sum_{c=1}^{C} e^{\vec{\mathbf{x}_\mathbf{i}} \cdot \vec{\beta_\mathbf{c}}}}$$

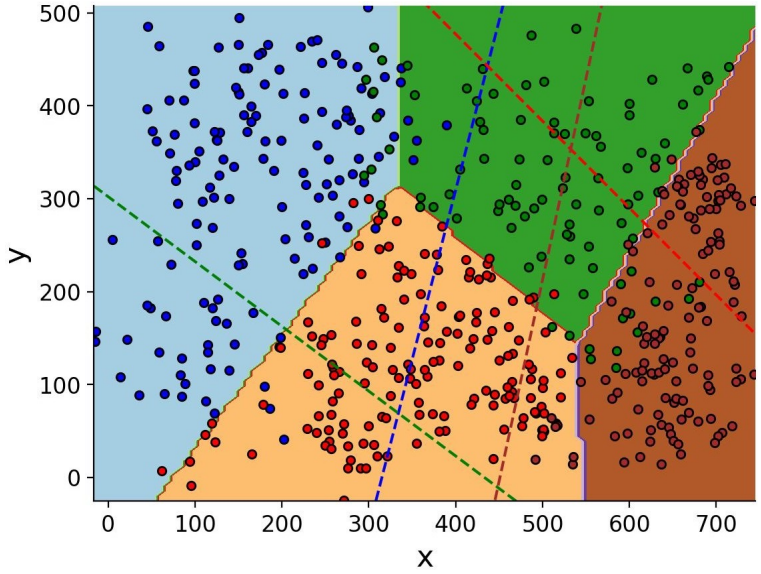
~ Ықтималдық үлгісі c’ класына жатады

Softmax функциясы



Көп класс классификациясы - OVR және көпмүшелік

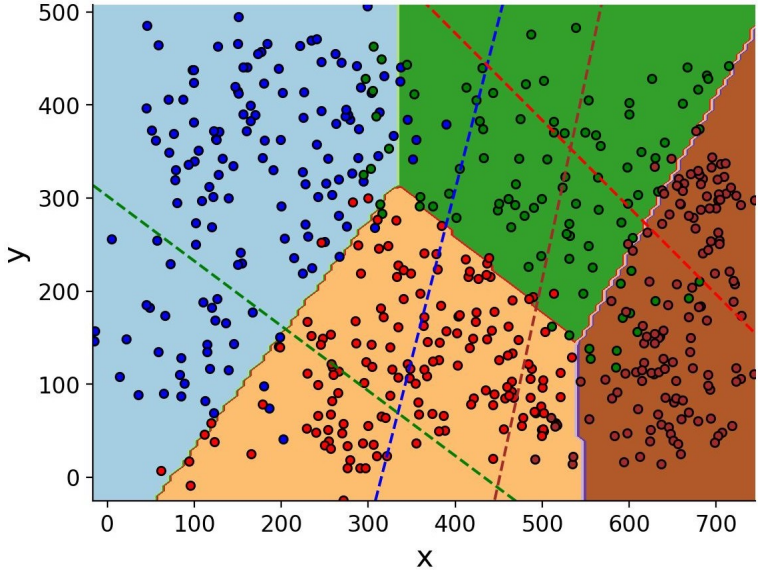




Көп кластық классификацияны бағалау

Confusion Matrix

Predicted Labels / True Labels	0	1	2	4
0	135	6	6	0
1	9	71	3	13
2	6	5	150	0
4	0	5	5	121



Көп кластық классификацияны бағалау

Дәлдік енді қиынырақ!

Екілік: 50/50 белгілер үшін кездейсоқ жіктеуіш 50% дәлдік алады

Multiclass: C белгілерінің бірдей саны үшін кездейсоқ жіктеуіш 1 / C дәлдігін алады 4 сынып үшін кездейсоқ жіктеуіш 25% дәлдік алады.

Көп кластық классификацияны бағалау

- Дәлдік, еске түсіру және F1 барлығының көп класс баламалары бар.
- Сіз оларды кез келген жеке сынып үшін есептей аласыз.

$$Precision_c = \frac{TP_c}{TP_c + FP_c}$$

- Барлық сыныптар бойынша есептеулер үшін:

Макроорташа: әрбір сынып үшін метриканы есептеп, содан кейін барлық сыныптар бойынша орташа мәнді алыңыз.

$$Precision = \frac{\sum_c TP_c}{\sum_c TP_c + FP_c}$$

  - Микроорташа: барлық сыныптар үшін алымдар мен бөлгіштерді есептеу.

$$Precision = \frac{\sum_c TP_c}{\sum_c TP_c + FP_c}$$