Naive Bayes (cont.)+ Linear Predictors

AW

Lecture Overview

Naive Bayes Estimates

Given a sample set S of size m where k samples are drawn from class $Y = 0$ and $m - k$ samples are drawn from class $Y = 1$. We showed that

- Under assumption of Bernoulli distribution fo class probabilities we have the max-log-likelyhood estimate for $\Pr(Y = y)$ is:

$$\Pr(Y = 1) = \frac{1}{m}\sum_{i=1}^{m} y_i \quad \text{and} \quad \Pr(Y = 0) = 1 - \frac{1}{m}\sum_{i=1}^{m} y_i$$

  wherey_i is the class value in i^{th} sample.

- For an atribute X that has finite domain $\{1, \ldots, d\}$ under comncentrated mass (pmf) assumption max-log-likelyhood estimate of conditional probability that attribute takes value i conditioned on class is:

$$\Pr(X = i|Y = 0) = \frac{x_i}{k} \quad \left(\text{resp. } \Pr(X = i|Y = 0) = \frac{x_i}{m - k}\right)$$

  where x_i is the number of times attribute X takes value i in subset of S that contains samples of class $Y = 0$ (resp. subset of S that contains samples of class $Y = 1$).

Naive Bayes Estimates

Given a sample set S of size m where k samples are drawn from class $Y = 0$ and $m - k$ samples are drawn from class $Y = 1$. We showed that

- Under assumption that conditioned on class real valued attribute X is normally distribuited max-log-likelyhood estimate of mean and standard deviation of this normal distribution are:

$$\hat{\mu} = \frac{\sum_{i=1}^{k} z_i}{k} \qquad \hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{k} (z_i - \mu)^2}{k}}$$

- Under the generative assumption (about probability distribution on data) that given a class, all features are independent of each other (i.e. $\Pr_{\mathscr{D}}(x_1 = a_1, \ldots, x_n = a_n | Y = y) = \prod_{i=1}^{n} \Pr(x_i = a_i | Y = y)$) for data coming from 2 random souurces (2 classes) the best predictor is naive Bayes classifier:

$$f_{\mathscr{D}}(x) = \begin{cases} 1 \text{ if } \frac{\Pr_{\mathscr{D}}(y=1|x)}{\Pr_{\mathscr{D}}(y=0|x)} = \frac{\Pr(Y=1) \prod_{i=1}^{n} \Pr(x_i=a_i|Y=1)}{\Pr(Y=0) \prod_{i=1}^{n} \Pr(x_i=a_i|Y=0)} \geq 1 \\ 0 \text{ otherwise} \end{cases}$$

Lecture Overview

Classification Example

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Evade (class) $\sim$ Bernoulli: let 'Yes'$= 1$, 'No'$= 0$, $m = |S| = 10$ so
$\Pr(\text{Yes}) = \Pr(Y = 1) = \frac{1}{m}\sum_{i=1}^{m} y_i = \frac{3}{10}$, and
$\Pr(\text{No}) = 1 - \Pr(\text{Yes}) = \frac{7}{10}$

Refund $\sim$ Bernoulli: let 'Yes'$= 1$, 'No'$= 0$.

For class Evade=Yes we have subsample
$S\,|_{\text{Evade=Yes}} = \{\text{rows } 5, 8, 10\}$, so $k = 3$ and
$\Pr(\text{Refund=Yes} \mid \text{Evade=Yes}) = \frac{1}{3}\sum_{i=1}^{k} y_i = \frac{0}{3}$, and
$\Pr(\text{Refund=No} \mid \text{Evade=Yes}) = 1 - \Pr(\text{Refund=Yes} \mid \text{Evade=Yes}) = 1 - 0 = 1$

Similarly for class Evade=No
$S\,|_{\text{Evade=No}} = \{\text{rows } 1, 2, 3, 6, 7, 9\}$ so $k = 7$ and
$\Pr(\text{Refund=No} \mid \text{Evade=No}) = \frac{4}{7}$ and
$\Pr(\text{Refund=No} \mid \text{Evade=No}) = \frac{3}{7}$

Classification Example

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Status has domain { married (=M), single (=s), divorced (=d) } .

For class Evade=Yes we have subsample $S\mid_{\text{Evade=Yes}} = \{\text{rows } 5, 8, 10\}$, so $k = 3$ and

$$\Pr(\text{Status=m} \mid \text{Evade=Yes}) = \frac{n_{\text{m} \wedge \text{yes}}}{k} = \frac{0}{3},$$

$$\Pr(\text{Status=d} \mid \text{Evade=Yes}) = \frac{n_{\text{d} \wedge \text{yes}}}{k} = \frac{1}{3},$$

$$\Pr(\text{Status=s} \mid \text{Evade=Yes}) = \frac{n_{\text{s} \wedge \text{yes}}}{k} = \frac{2}{3}.$$

For class Evade=Yes we have subsample $S\mid_{\text{Evade=No}} = \{\text{rows } 1, 2, 3, 4, 6, 7, 9\}$, so $k = 7$ and

$$\Pr(\text{Status=m} \mid \text{Evade=No}) = \frac{4}{7},$$

$$\Pr(\text{Status=d} \mid \text{Evade=No}) = \frac{1}{7},$$

$$\Pr(\text{Status=s} \mid \text{Evade=Yes}) = \frac{2}{7}.$$

Assuming conditioned on 'Evade' class income feature is distributed normally.

For class Evade=yes we have $\mu_{i,\text{yes}} = \frac{95+85+90}{3} = 90$ and

$\sigma_{i,\text{yes}} = \sqrt{\frac{(95-90)^2+(85-90)^2+(90-90)^2}{3}} \approx 4.82$ and

For class Evade=No we have $\mu_{i,\text{no}} = \frac{125+100+70+120+60+220+75}{7} = 110$ and $\sigma_{i,\text{no}} =$

$$\sqrt{\frac{(125-110)^2+(100-110)^2+(70-110)^2+(120-110)^2+(60-110)^2+(220-110)^2+(75-110)^2}{7}} \approx 50.5$$

Classification Example - continued

Given a test data point (Refund=No, Status=Divorced, Income=120), what is its 'Evade' class?

$$\begin{aligned}\Pr((\text{R=No,S=d,I=120})|\text{E=Y})\Pr(\text{E=Y}) &= \Pr(\text{R=No}|\text{E=Y})\Pr(\text{S=d}|\text{E=Y})\Pr(\text{I=120}|\text{E=Y})\Pr(\text{E=Y}) \\ &= 1 \times \tfrac{1}{3} \times \tfrac{1}{4.82\sqrt{2\pi}} \exp\left(-\tfrac{(120-90)^2}{2\cdot 4.82^2}\right) \times 0.3 \\ &\approx 3.2 \times 10^{-11}\end{aligned}$$

$$\begin{aligned}\Pr((\text{R=No,S=d,I=120})|\text{E=N})\Pr(\text{E=N}) &= \Pr(\text{R=No}|\text{E=N})\Pr(\text{S=d}|\text{E=N})\Pr(\text{I=120}|\text{E=N})\Pr(\text{E=N}) \\ &= \tfrac{4}{7} \times \tfrac{1}{7} \times \tfrac{1}{50.5\sqrt{2\pi}} \exp\left(-\tfrac{(120-110)^2}{2\cdot 50.5^2}\right) \times 0.7 \\ &\approx 0.00044\end{aligned}$$

Since

$$\Pr((\text{R=No,S=d,I=120})|\text{E=N})\Pr(\text{E=N}) = 0.00044 > 3.2\times 10^{-11} = \Pr((\text{R=No,S=d,I=120})|\text{E=Y})\Pr(\text{E=Y})$$

we conclude that the 'Evade' class of the data point is No.

Other Models of Features (1 of 3)

What if our continuous feature is **not normally distributed?** For example what what if the feature is the amount of time that was spent on a site before purchase? This is know to have *exponential distribution* that is 0 for $x \leq 0$ and for $x > 0$ with cdf/pdf respectively

$$F(x;\lambda) = 1 - e^{-\lambda x} \quad \text{and} \quad f(x;\lambda) = \lambda e^{-\lambda x}$$

We have sample $S = \{x_1, \ldots, x_n\}$ i.i.d. drawn from $F(x;\lambda)$. Parameter of estimate is λ, so its likelihood is

$$L(\lambda; x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i;\lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n e^{-\sum_{i=1}^{n} \lambda x_i}$$

and $\log$-likelihood is $l(\lambda; x_1, \ldots, x_n) = n \ln \lambda - \sum_{i=1}^{n} \lambda x_i$. So

$$\frac{d}{d\lambda} l(\lambda; x_1, \ldots, x_n) = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i = 0 \quad \text{yields} \quad \lambda = \frac{n}{\sum_{i=1}^{n} x_i}$$

i.e. reciprocal of average, not the average!

Other Models of Features (2 of 3)

What if the feature is the bid for certain type of goods that is sold on many-times repeated auctions? This is know to have *uniform distribution* between low value a and high value b that is 0 outside the range and within the range has pdf $f(x; a, b) = \frac{1}{b-a}$.
We have sample $S = \{x_1, \ldots, x_n\}$ i.i.d. drawn from $f(x; \lambda)$. Parameters of estimate are a and b, so its likelihood is

$$L(a, b; x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i; a, b) = \prod_{i=1}^{n} \frac{1}{b-a} = \frac{1}{(b-a)^n}$$

and $\log$-likelihood is $l(a, b; x_1, \ldots, x_n) = -n \ln (b - a$. So the derivative wrt a is $\frac{d}{d\,a} l(a, b; x_1, \ldots, x_n) = \frac{n}{b-a}$ and with respect to b is $\frac{d}{d\,b} l(a, b; x_1, \ldots, x_n) = -\frac{n}{b-a}$.
Notice that the derivative with respect to a is monotonically increasing. Thus, the mle for a would be the smallest value of a possible, which would simply be $\min_{x_i \in S}\{x_i\}$. Similarly the derivative with respect to b is monotonically decreasing. Thus, the mle for b would be the largest b possible which would simply be $\max_{x_i \in S}\{x_i\}$.

Other Models of Features(3 of 3)

- We assumed direct pmf model of finite domains. The problem: conditional probability of one of the features could be zero, e.g. $\Pr(\text{Status=Married}|\ \text{Evade=Yes}) = 0$. Then the entire expression for the class (in our case Evade=Yes) becomes zero.
  - it is just because of lack of data - shouldn't discount the class!
- Solution: instead of log-likelihood estimate of direct pmf model of finite domains use maximum a posteriori (MAP) estimate of direct pmf model of finite domains. It requires knowledge of prior probability distribution. What is MAP and why prior is needed?

Other Models of Features(3 of 3)

- Solution: instead of log-likelihood estimate of direct pmf model of finite domains use maximum a posteriori (MAP) estimate of direct pmf model of finite domains. It requires knowledge of prior probability distribution. What is MAP and why prior is needed?
- Let prior distribution g over parameter value θ is known. This allows us to treat θ as a random variable. We can calculate the posterior distribution of θ using Bayes' theorem:

$$P(\theta \mid x = a) = \frac{P(x = a|\theta)g(\theta)}{\int_\Theta P(x = a|\theta)g(\theta)d\theta}$$

  where g is density function of θ and Θ is the domain of θ.
- The MAP estimation then estimates θ as the mode of the posterior distribution of this random variable:

$$\begin{aligned}\theta_{\mathrm{MAP}}(x) &= \arg\max_\theta P(\theta \mid x = a) = \arg\max_\theta \frac{P(x=a|\theta)g(\theta)}{\int_\Theta P(x=a|\theta)g(\theta)d\theta} \\ &= \arg\max_\theta P(x = a|\theta)g(\theta)\end{aligned}$$

Model for Features with Finite Domains

- Suppose a probabilistic experiment can have only two outcomes, either success (feature$= i$), with probability x, or failure (feature$\neq i$), with probability $1 - x$. Suppose also that x is unknown and all its possible values are deemed equally likely. This uncertainty can be described by assigning to x a uniform distribution on the interval $[0, 1]$.
- Suppose that we perform n independent repetitions of the experiment and we observe k successes and $n - k$ failures. After performing the experiments, we naturally want to know how we should revise the distribution initially assigned to x. In other words, we want to calculate the conditional distribution of x, conditional on the number of successes and failures we have observed.
- Provable: this conditional distribution of x is a Beta distribution with parameters $k + 2$ and $n - k + 2$.

Model for Features with Finite Domains

- Density of Beta probability distribution $\beta(\theta, \alpha, \beta)$ is given by $\beta(\theta, \alpha, \beta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{\int_0^1 u^{\alpha-1}(1-u)^{\beta-1}du}$. Thus in case when we observe k success cases (i.e. when feature $= i$) and $n-k$ failure cases (i.e. when feature $\neq i$) we have $\beta(\theta, k+2, n-k+2) = \frac{\theta^{k+1}(1-\theta)^{n-k+1}}{\int_0^1 u^{k+1}(1-u)^{n-k+1}du}$.
- Then MAP probability estimate for $\beta(\theta, \alpha, \beta)$ is $\theta_{MAP}(\alpha, \beta) = \frac{\alpha-1}{\beta+\alpha-2}$
  - in particular under $\alpha = k$ and $\beta = n-k$ we get Laplace probability estimate $\theta_{Laplace} = \frac{k+1}{n+2}$

The Naive Bayesian in R - e1071

The standard naive Bayes classifier assumes

- independence of the predictor variables,
- Gaussian distribution (given the target class) of rel-valued features
- For attributes with missing values, the corresponding table entries are omitted for prediction

Returns

- a priori probabilities for each class
- Conditional probabilities for each feature
- Feature means and variances for each class for each real-valued feature

Iris Example in **R**

```
library(e1071)
data("iris")
set.seed(2)
y<-sample(1:nrow(iris),2*nrow(iris)/3,F)
nBayes <- naiveBayes(Species ~ .,
                     data = iris[y,]);nBayes
table(iris[y,]$Species,
                predict(nBayes,iris[y,-5]))
table(iris[-y,]$Species,
          predict(nBayes,newdata=iris[-y,-5]))
```

Reading

SSBD sections 24.1, 24.2
You can skip proofs if you are not interested in technicalities

TSK sections 5.3.1, 5.3.2, 5.3.3

Lecture Overview

Flat, Translate, Linear Functional

Definition

i. A translate of a set s in $\mathbb{R}^n$ by a vector $\overline{p}$ is the set $S + \overline{p} = \{\overline{x} + \overline{p} \mid \overline{x} \in S\}$

ii. A flat in $\mathbb{R}^n$ is a translate of a subspace of $\mathbb{R}^n$. Two flats are parallel if one is a translate of the other

iii. The dimension of a flat is the dimension of the corresponding parallel subspace. The dimension of a set S, written as $\dim S$, is the dimension of the smallest flat containing S

Note that a line in $\mathbb{R}^n$ is a flat of $\dim = 1$. A hyper-plane in $\mathbb{R}^n$ is a translate of a plane, i.e. a flat of dimension $\dim = n - 1$

Definition

A linear functional on $\mathbb{R}^n$ is a linear transformation f from $\mathbb{R}^n$ to R. For each scalar d in $\mathbb{R}$, the symbol $[f : d]$ denotes the set of all vectors $\overline{x}$ in $\mathbb{R}^n$ at which the value of $f(\overline{x})$ is d. That is, $[f : d] = \{\overline{x} \in \mathbb{R}^n \mid f(\overline{x}) = d\}$.

Hyperplanes

- Note that linear functional $f : \mathbb{R}^n \to \mathbb{R}$ is given by a row-matrix (=vector transpose) $[w_1, \ldots, w_n]$, i.e. $f(\overline{x}) = \overline{w} \bullet \overline{x}$.
- The zero functional f_0 is such linear functional that $f_0(\overline{x}) = 0$ for all $\overline{x} \in \mathbb{R}^n$. Other functionals are said to be non-zero.
- If f is a linear functional on $\mathbb{R}^n$, with the standard matrix $\overline{w}^T$ then

$$[f : 0] = \{\overline{x} \in \mathbb{R}^n \mid \overline{w} \bullet \overline{x} = 0\} = \mathbf{Nul}(\overline{w}^T).$$

  If $\overline{w}$ is a nonzero, then rank $\overline{w}^T = 1$ so $\dim \mathbf{Nul}(\overline{w}^T) = n - 1$ by the rank/nullity theorem. So, $\dim[f : 0] = n - 1$ hence $[f : 0]$ is a hyper-plane.
- For $d \in \mathbb{R}$, $[f : d] = \{\overline{x} \in \mathbb{R}^n \mid \overline{w} \bullet \overline{x} = d\} = \overline{p} + [f : 0]$ where $\overline{p}$ is a particular solution of nonhomogeneous equation $\overline{w}^T \overline{x} = d$. So $[f : d]$ is a hyper-plane parallel to $[f : 0]$ translated through $\overline{p}$.

Can all hyperplanes in $\mathbb{R}^n$ be described as $[f : d]$ for some f and d?

Set Separation by Hyperplanes

Theorem

A subset H of $\mathbb{R}^n$ is a hyper-plane if and only if $H = [f : d]$ for some nonzero linear functional f and some real d.

Thus, if H is a hyperplane, there exist a nonzero vector $\overline{w}$ and a real number d such that $H = \{\overline{x} | \overline{w} \bullet \overline{x} = d\}$. It is sometimes more convenient to define H by a vector $\overline{w'} = (d, w_1, \ldots, w_n)^T$ and then
$H = \{\overline{x'} \in \mathbb{R}^{n+1} \mid ovx' = (1, x1, \ldots, x_n)^T$ and $\overline{w'} \bullet \overline{x'} = 0\}$

Definition

The hyper-plane $H = [f : d]$ separates two sets A and B if one of the following holds:

i. $f(A) \leq d$ and $f(B) > d$, or

ii. $f(A) \geq d$ and $f(B) < d$.

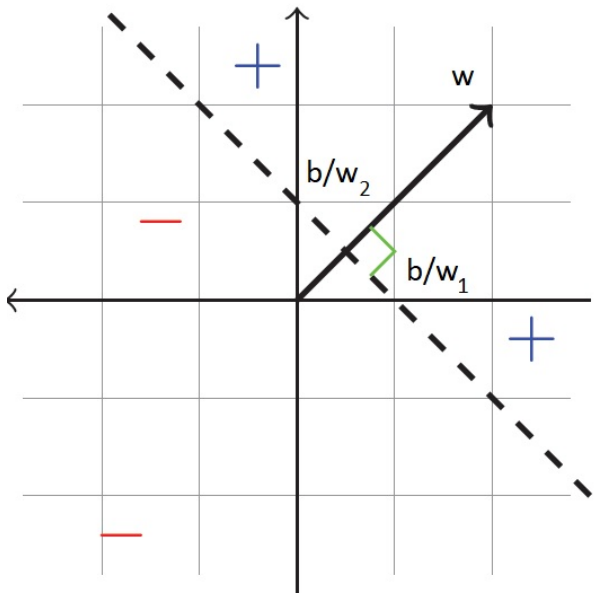
If in the conditions above all the weak inequalities are replaced by strict inequalities, then H is said to strictly separate A and B.

Halfspaces

The class of halfspace classifiers separate instances using a family of hyperplanes $\mathbb{L} = \{L_n\}_{n=1}^{\infty}$ where $L_n = \{H = [f : d] \mid f \equiv \overline{w} \in \mathbb{R}^n, d \in R\}$ to separate classes, i.e. hyperplane classifier $H = [f : d]$ where $f \equiv \overline{w} \in \mathbb{R}^n$ computes class y for an instance $\overline{x} \in \mathbb{R}^n$ as $y = \mathbf{sign}(\overline{w} \bullet \overline{x} - d)$.

Solving hyperplane equation $\overline{w} \bullet \overline{x} = d$ gives that hyperplane intercepts with i^{th} axis at $\frac{d}{w_i}$.

If we are looking for a linear predictor that is ERM predictor w.r.t realizable PAC learning case, then for a given sample set $S = \{(\overline{x}_1, y_1), \ldots, (\overline{x}_m, y_m)\}$ we need to find $\overline{w}$ and $d \in \mathbb{R}$ such that for every $i \in \{1, \ldots, m\}$ we have $\mathbf{sign}(\overline{w} \bullet \overline{x}_i - d) = y_i$.
Equivalently $y_i \cdot (\overline{w} \bullet \overline{x}_i - d) > 0, \quad \forall i = 1, \ldots, m$

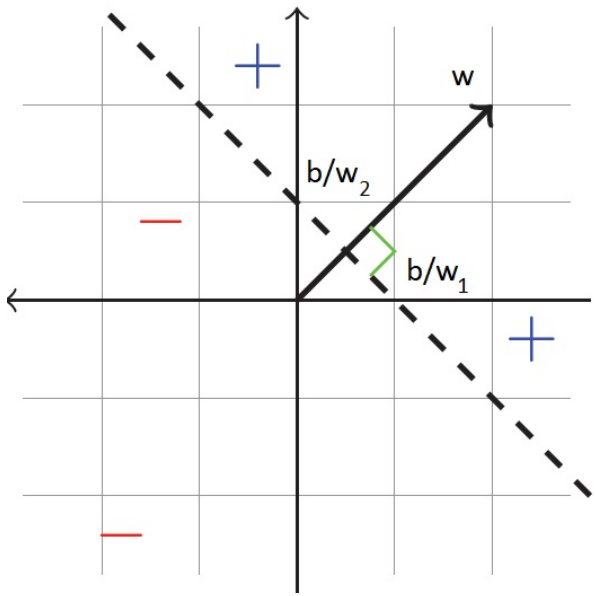


There must be a solution for the system of m inequalities $y_i \cdot (\overline{w} \bullet \overline{x}_i - d) > 0, \quad \forall i = 1, \ldots, m$ because we assume that it is realizable case. To solve we could use Linear programming, but inequalities are strict so LP is inapplicable.

Halfspaces

The class of halfspace classifiers separate instances using a family of hyperplanes $\mathbb{L} = \{L_n\}_{n=1}^{\infty}$ where $L_n = \{H = [f : d] \mid f \equiv \overline{w} \in \mathbb{R}^n, d \in R\}$ to separate classes, i.e. hyperplane classifier $H = [f : d]$ where $f \equiv \overline{w} \in \mathbb{R}^n$ computes class y for an instance $\overline{x} \in \mathbb{R}^n$ as $y = \mathbf{sign}(\overline{w} \bullet \overline{x} - d)$.

Solving hyperplane equation $\overline{w} \bullet \overline{x} = d$ gives that hyperplane intercepts with i^{th} axis at $\frac{d}{w_i}$.



If we are looking for a linear predictor that is ERM predictor w.r.t realizable PAC learning case, then for a given sample set $S = \{(\overline{x}_1, y_1), \ldots, (\overline{x}_m, y_m)\}$ we need to find $\overline{w}$ and $d \in \mathbb{R}$ such that for every $i \in \{1, \ldots, m\}$ we have $\mathbf{sign}(\overline{w} \bullet \overline{x}_i - d) = y_i$. Equivalently $y_i \cdot (\overline{w} \bullet \overline{x}_i - d) > 0, \quad \forall i = 1, \ldots, m$

To find equivalent system with $\geq$ constraints, suppose $\overline{w}^*, d^*$ is a solution. Let then $\gamma = \min_i(y_i(\overline{w}^* \bullet \overline{x}_i - d^*))$ and let $\overline{w}^\dagger = \frac{\overline{w}^*}{\gamma}$ and $d^\dagger = \frac{d^*}{\gamma}$. Then $\frac{1}{\gamma} y_i \cdot (\overline{w}^* \bullet \overline{x}_i - d^*) = y_i \cdot (\overline{w}^\dagger \bullet \overline{x}_i - d^\dagger) \geq 1, \quad \forall i = 1, \ldots, m$. Vector satisfying these conditions can be found by solving LP with dummy objective $(\min 0)$.