Posplošeni linearni modeli

Edini napovedni model, ki smo si ga ogledali do sedaj, je linearna regresija. Ponovimo: na vhodu imamo značilke z zveznimi vrednostmi x_i, ki jih utežimo z utežmi θ_i ter jim prištejemo prosti člen (intercept) θ_0, s čimer dobimo napoved ciljne vrednosti $\hat{y}$. Gradnja napovednega modela poteka tako, da iščemo takšne vrednosti parametrov, ki minimizirajo vsoto kvadratov napak med dejanskimi in napovedanimi vrednostmi ciljne spremenljivke v učni množici. Linearna regresija je eden najenostavnejših regresijskih modelov; enostavnejši pristop bi bil le napovedovanje s povprečno vrednostjo odvisne spremenljivke v učni množici, vendar tak pristop ne upošteva vrednosti vhodnih značilk, zato ga težko obravnavamo kot pravi napovedni model

V tem poglavju nam bo koristil vektorski in matrični zapis zgornjega modela, kjer podatke zapišemo v matriko značilk X, parametre združimo v vektor $\boldsymbol{\theta}$, napovedi pa izrazimo kot $\hat{\mathbf{y}} = X\boldsymbol{\theta}$, kar omogoča kompaktnejši zapis in učinkovitejšo implementacijo algoritmov za učenje modela. Pri tem predpostavimo, da je prva kolona v matriki podatkov kar enotski vektor, da sicer ne kompliciramo s posebno obravnavo s prostim členom in da gredo indeksi za θ od 0 do d, kjer je d število neodvisnih spremenljivk.

Postavi se vprašanje, ali obstajajo drugi, podobno enostavni modeli, morda za nekoliko drugačne napovedne naloge, torej takšni, kjer bi značilke ravno tako povezali z uteženo vsoto, nato pa dobljeno vrednost preoblikovali z ustrezno (nelinearno) povezovalno funkcijo, tako da bi lahko modelirali tudi diskretne ali kako drugače omejene ciljne spremenljivke. V nadaljevanju začnemo s primerom takšnega modela, ki ga bomo uporabili za razvrščanje, nato pa se vprašamo, ali so tovrstne razširitve dovolj splošne za širši razred modelov, kaj pri njih pravzaprav predpostavimo, od kod izhajajo njihove kriterijske funkcije in ali gre pri tem za zanimiv razred modelov s skupnimi lastnostmi.

Primer

Začnimo z (izmišljenim) primerom. V tabeli 12 so zbrani kopalci na Bledu, ki smo jih vprašali, koliko ur na teden se ukvarjajo s športom in koliko ur so prejšnjo noč spali. Zabeležili smo tudi, ali so v dnevu intervjuvanja uspeli odplavati na otok. Ta je od bližnjega kopališča oddaljen več kot pol kilometra v eni smeri, zato je plavanje na otok in nazaj kar zalogaj, ki ne bi bil ravno primeren za slabše kopalce. Cilj je razviti aplikacijo, ki bi kopalcem svetovala, seveda glede na fizično pripravljenost in spočitost, ali naj se napotijo na tak podvig. Aplikacija seveda potrebuje napovedni model, tega pa lahko zgradimo iz naših podatkov.

Slika 12: Primer klasifikacijskih podatkov z meta atributum, neodvisnima spremenljivkama in razredom (Otok).

Ime	Vadba	Spanje	Otok
Alenka	7	8	1
Ana	2	8	0
Andrej	5	5	0
Blaž	5	9	1
Boštjan	7	5	0
Goran	12	6	1
Gregor	10	9	1
Helena	4	9	1
Irena	9	3	0
Janez	5	6	0
Jure	8	4	0
Katarina	4	3	0
Klara	3	9	1
Luka	5	4	0
Maja	9	6	0
Marko	4	7	0
Matej	4	8	1
Miha	6	4	0
Mojca	11	5	1
Nika	2	5	0
Nina	8	6	1
Petra	3	6	0
Polona	9	8	1
Rok	10	6	1
Sara	6	7	1
Sašo	10	7	1
Sebastjan	12	5	1
Tatjana	12	9	1
Tjaša	11	3	0
Tomaz	12	5	1

Ker so podatki dvodimenzionalni, jih je najbolje izrisati v razsevnem diagramu. Označimo tudi razred. Že prvi pogled na izris kaže, da je morda mogoče dobre plavalce in tiste, ki se bolj kopajo, ločiti s črto oziroma odločitveno mejo. Ta je linearna, zato jo lahko zapišemo kot $X\theta = 0$. Izris vključuje tudi tri nove obiskovalce: Saro, Martina in Leona. Kateremu izmed njih bo naša aplikacija oziroma model svetovala, da lahko odplava do otoka in priplava nazaj?

Sara je na strani plavalcev. Je daleč od odločitvene meje ki loči med obema razredoma. Prav gotovo lahko plava do otoka in nazaj. Martin je zelo na drugi strani, nikakor naj se ne oddalji od obale. Leon je, kar se tiče odločitvene meje, na strani plavalcev, a za las. Svetovati mu da naj poskusi plavati do otoka bi bilo zelo narobe. Naš problem je sicer klasifikacijski, želimo napovedati enega od dveh možnih razredov, a bolje bi bilo to narediti previdno, z uporabo verjetnosti. Ker je Sara zelo oddaljena od odločitvene meje, je prav gotovo kandidatka za odhod na otok, Martin nikakor ne, Leon pa je nekje na maje, njegova verjetnos "plavalnega" razreda je okoli 50%.

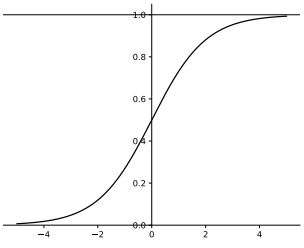
Postaja nam jasno: oddaljenost od odločitvene meje moramo pretvoriti v verjetnosti. Kar seveda ne bo problem. Linearna kombinacija $z = X\theta$ je pravzaprav proporcionalna oddaljenosti od premice, ki jo določajo parametri θ. Za pretvorbo lahko uporabimo funkcijo, katere zaloga vrednosti je med 0 in 1. Primer take funkcije je sigmoida $\sigma(z)$, naša verjetnost pa je potem

$$P(y = 1 \mid X) = \sigma(z) = \frac{1}{1 + e^{-z}}.$$

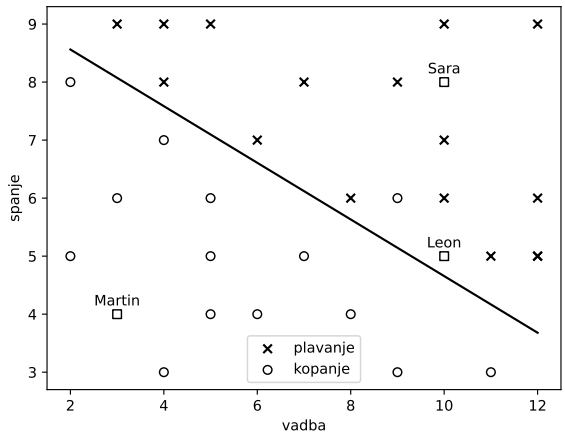
Odločitvena meja s slike ima parametre parametre $\theta_0 = -15.6$, $\theta_1 = 0.8$ in $\theta_2 = 1.6$, zato lahko odločitveno enačbo za oddaljenost od te meje zapišemo kot

$$z = -15.6 + 0.8 \cdot \text{vadba} + 1.6 \cdot \text{spanje}.$$

Za nove primere dobimo: Sara ima $z = 5.5$ in $P(\text{otok} = 1) \approx 1.0$, zato je zelo primerna kandidatka za plavanje do otoka; Martin ima



Slika 13: Sigmoidna funkcija.



Slika 14: Učni podatki, možna ločitvena meja med razredoma, in novi (imenovani) primeri, ki jih moramo še razvrstiti.

$z = -6.7$ in $P(\text{otok} = 1) \approx 0.0$, zato mu to odsvetujemo; Leon pa ima $z = 0.6$ in $P(\text{otok} = 1) \approx 0.6$, kar pomeni, da je blizu odločitvene meje in je odločitev precej negotova.

Model, ki smo ga precej na hitro in morda malo površno uvedli na našem primeru se imenuje logistična regresija. Pomembno je, da tako kot linearna regresija tudi ta model uporablja linearno kombinacijo neodvisnih spremenljivk, katere rezultat pa tokrat, čisto zato, da lahko vrnemo verjetnosti, transformiramo z sigmoidno funkcijo. A postavlja se vprašanje: kako se sploh naučimo "pravih" parametrov našega modela, toej vektorja θ? Kakšno kriterijsko funkcijo za to optimiziramo? Iz kakšnih predpostavk ta izhaja? Poznamo poleg linearne in logistične regresije še kakšne druge modele te vrste? In končno, je tudi to moč uporabiti strojno odvajanje in gradientni sestop za učenje modela?

Čas je za malce teorije.

Eksponentna družina porazdelitev

Od kod torej izhajajo modeli, kot sta linearna in logistična regresija? Kakšne predpostavke sploh pri tem naredimo o podatkih? Uberemo podoben pristop kot ga že poznamo pri linearni regresiji: namesto, da bi si izmislili kriterijsko funkcijo (npr. vsota kvadratov napak na učni množici), bomo izhajali iz verjetnostnega modela, torej modela, ki generira podatke v učno množico, in iz te predpostavke izpeljali kriterijsko funkcijo.

Razred porazdelitev, ki se za naš namen izkaže še posebej uporaben, je *eksponentna družina*. Porazdelitev spada v to družino, če njeno verjetnostno funkcijo (ali gostoto) za spremenljivko y lahko zapišemo v

obliki

$$p(y \mid \theta) = h(y) \exp\left(\eta(\theta)\, T(y) - A(\theta)\right),$$

kjer θ označuje parameter porazdelitve. Ker nas bo pri strojnem učenju zanimal logaritem verjetja, je smiselno ta izraz logaritmirati:

$$\log p(y \mid \theta) = \eta(\theta) T(y) - A(\theta) + \log h(y).$$

Funkciji $T(y)$ pravimo *zadostna statistika* in je v najpreprostejših primerih kar enaka y. Funkcija $\eta(\theta)$ je t. i. *naravni parameter*, ki predstavlja preoblikovanje osnovnega parametra θ. Funkcija $A(\theta)$ skrbi za normalizacijo porazdelitve (ta faktor skrbi, da se verjetnosti seštejejo oziroma integrirajo v 1), $h(y)$ pa je od parametra neodvisen del. Opazimo, da je logaritemska verjetnost linearna v $T(y)$, kar omogoča enostavno optimizacijo in povezavo z linearnimi modeli.

V eksponentni družini porazdelitev najdemo večino porazdelitev, ki nas zanimajo v strojnem učenju in ki so povezane z regresijo (napovedovanje zveznega razreda), klasifikacijo (napovedovanje diskretne spremenljivke) in modeliranjem štetij. Te porazdelitve so naprimer normalna, Bernoullijeva in Poissonova. A podrobnosti o tem seveda sledijo.

Naravni parameter in povezovalna funkcija

V nadzorovanem učenju želimo modelirati pogojno pričakovano vrednost ciljne spremenljivke, to je njeno povprečno vrednost pri danih vhodnih značilkah x (v primeru klasifikacije pa to ustreza verjetnosti pripadnosti posameznemu razredu):

$$\mu(x) = \mathbb{E}[y \mid x].$$

Pri posplošenih linearnih modelih naredimo ključno predpostavko, da je naravni parameter povezan z značilkami prek linearnega napovednika

$$\eta = X\theta.$$

To ni posledica eksponentne družine, temveč modelna predpostavka, s katero razširimo idejo linearne regresije na širši razred porazdelitev. Za porazdelitve iz eksponentne družine pa velja, da je pričakovana vrednost določena z naravnim parametrom, $\mu = A'(\eta)$. V tipičnih primerih lahko to zvezo obrnemo in zapišemo

$$\eta = g(\mu).$$

S tem dobimo model

$$g(\mu(x)) = X\theta,$$

kjer funkciji g pravimo povezovalna funkcija.

Ker se mora gostota normirati, velja

$$\int h(y) \exp\left(\eta T(y) - A(\eta)\right) dy = 1.$$

Odvajanje po η da

$$\mathbb{E}[T(y)] - A'(\eta) = 0,$$

zato

$$\mathbb{E}[T(y)] = A'(\eta).$$

Če funkcijo g izberemo tako, da sovpada z naravno zvezo med μ in η, govorimo o *kanončni povezovalni funkciji*. Ta izbira vodi do posebej enostavnih izrazov za verjetje in njegove odvode ter pogosto do konveksnih optimizacijskih problemov.

Učenje modela

Tule samo spomnimo, da bomo za določitev parametrov modela potrebovali kriterijsko funkcijo, za to pa potrebujemo verjetje oziroma njegov logaritem. Pravzaprav imamo za to že vse pripravljeno. Predpostavimo, da so učni primeri med seboj neodvisni, in ko se še odločimo za ciljno porazdelitev razredov, lahko zapišemo verjetje za učne podatke

$$p(\mathbf{y} \mid X, \theta) = \prod_{i=1}^{n} p(y_i \mid x_i, \theta).$$

Za učenje parametrov maksimiziramo logaritemsko verjetje

$$\ell(\theta) = \sum_{i=1}^{n} \log p(y_i \mid x_i, \theta),$$

kar je ekvivalentno minimizaciji negativnega log-verjetja, ki ga lahko razumemo kot kriterijsko funkcijo.

Na tej točki postane povezava s strojnim učenjem očitna: različne izbire porazdelitev vodijo do različnih kriterijskih funkcij, optimizacija pa poteka enako, npr. z gradientnim sestopom.

Linearna regresija

Začnimo z najpreprostejšim primerom, ki smo ga že spoznali, a ga zdaj pogledamo z nove perspektive. Predpostavimo, da je ciljna spremenljivka zvezna in sledi normalni porazdelitvi

$$y \mid x \sim \mathcal{N}(\mu(x), \sigma^2).$$

Gostoto lahko zapišemo kot

$$p(y \mid x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right).$$

Če izraz preuredimo, dobimo

$$p(y \mid x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{\mu}{\sigma^2} y - \frac{\mu^2}{2\sigma^2} - \frac{y^2}{2\sigma^2}\right),$$

kar je oblike eksponentne družine

$$p(y \mid \theta) = h(y)\, \exp\big(\eta(\theta)\, T(y) - A(\theta)\big),$$

kjer prepoznamo:

$$T(y) = y, \qquad \eta(\theta) = \frac{\mu}{\sigma^2}, \qquad A(\theta) = \frac{\mu^2}{2\sigma^2}, \qquad \log h(y) = -\frac{y^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2).$$

Če predpostavimo, da je varianca σ^2 konstantna, je naravni parameter η sorazmeren z μ, zato ga lahko (z rahlo zlorabo zapisa) obravnavamo kar kot μ, ta pa je enak linearni kombinaciji značilk. V tem primeru je kanonična povezovalna funkcija identiteta,

$$g(\mu) = \mu = X\theta.$$

Logaritemsko verjetje za en primer je

$$\log p(y \mid x, \theta) = -\frac{1}{2\sigma^2}(y - X\theta)^2 + C,$$

kjer C ne zavisi od θ. Maksimizacija verjetja je zato ekvivalentna minimizaciji vsote kvadratov napak:

$$L(\theta) = \sum_{i=1}^{n}(y_i - x_i^T\theta)^2.$$

Kot že vemo, kriterijska funkcija linearne regresije neposredno izhaja iz predpostavke o normalni porazdelitvi šuma.

Logistična regresija

V primeru klasifikacije z dvema razredoma predpostavimo, da ciljna spremenljivka sledi Bernoullijevi porazdelitvi:

$$y \mid x \sim \text{Bernoulli}(p(x)),$$

kjer je x vektor opaženih značilk za posamezen primer, $y \in \{0, 1\}$ pa razred. Funkcija $p(x)$ predstavlja verjetnost, da je razred enak 1 glede na značilke x, torej

$$P(y = 1 \mid x) = p(x), \qquad P(y = 0 \mid x) = 1 - p(x).$$

Pričakovana vrednost ciljne spremenljivke y pri danih značilkah x je torej

$$\mu(x) = \mathbb{E}[y \mid x] = p(x).$$

Verjetnostno funkcijo ciljne spremenljivke y pri danih značilkah x lahko zapišemo v eni enačbi kot

$$p(y \mid x) = p^y(1 - p)^{1-y},$$

ter njen logaritem kot

$$\log p(y \mid x) = y \log p + (1 - y)\log(1 - p).$$

Ta verjetnost nam bo služila za zapis logaritemskega verjetja oziroma negativnega log verjetja.

Ta izraz preuredimo:

$$\log p(y \mid x) = y \log \frac{p}{1-p} + \log(1-p).$$

Pri predpostavki neodvisnosti učnih primerov je ta zapis že primeren za oblikovanje verjetja, ta pa za oblikovanje kriterijske funkcije. A to še sledi. Najprej moramo odgovoriti na problem, kako p povežemo z $X\theta$.

Zdaj lahko izraz primerjamo s splošno obliko eksponentne družine

$$\log p(y \mid \theta) = \eta(\theta)\, T(y) - A(\theta) + \log h(y),$$

in prepoznamo posamezne dele:

$$T(y) = y, \qquad \eta = \log \frac{p}{1-p}, \qquad A(\eta) = -\log(1-p), \qquad h(y) = 1.$$

Spremenljivka η, oziroma funkcija $\eta(p)$ (naravni parameter) preoblikuje verjetno p v logaritemsko razmerje obetov, torej v $\log \frac{p}{1-p}$ (angl. *log-odds*). Ker je $\mu = p$, dobimo povezovalno funkcijo, ki ji pravimo logit:

$$g(\mu) = \log \frac{p(x)}{1-p(x)} = X\theta,$$

kar vodi do znane oblike sigmoide, s katero smo, nekako intuitivno, začeli to poglavje:

$$p(x) = \frac{1}{1+e^{-X\theta}}.$$

Poissonova regresija

Za modeliranje štetij (npr. število dogodkov v nekem časovnem intervalu) predpostavimo Poissonovo porazdelitev:

$$y \mid x \sim \text{Poisson}(\lambda(x)).$$

To pomeni, da je $y \in \{0, 1, 2, \dots\}$ in velja

$$p(y \mid x) = \frac{\lambda^y e^{-\lambda}}{y!}.$$

Pričakovana vrednost te porazdelitve je

$$\mu(x) = \mathbb{E}[y \mid x] = \lambda(x).$$

Da vidimo povezavo z eksponentno družino, vzamemo logaritem:

$$\log p(y \mid x) = y \log \lambda - \lambda - \log(y!).$$

Ta izraz je že skoraj v obliki

$$\log p(y \mid x) = \eta\, y - A(\eta) + \log h(y),$$

od koder prepoznamo

$$\eta = \log \lambda.$$

Ker je $\mu = \lambda$, dobimo povezovalno funkcijo

$$g(\mu) = \log \mu,$$

ki ji pravimo logaritemska povezava. Linearni model tako zapišemo kot

$$\log \lambda(x) = X\theta,$$

od koder sledi

$$\lambda(x) = e^{X\theta}.$$

Logaritemsko verjetje za en primer je

$$\log p(y \mid x, \theta) = yX\theta - e^{X\theta} - \log(y!),$$

in negativno log-verjetje vodi do kriterijske funkcije

$$L(\theta) = \sum_{i=1}^{n} \left(e^{x_i^T \theta} - y_i x_i^T \theta \right),$$

pri čemer lahko člen $\log(y!)$ izpustimo, saj ne zavisi od parametrov.

Premor s kodiranjem

Tole zgoraj je bilo veliko teorije in malo primerov. Za en premor spišimo preverimo, ali z logistično regresijo res dobimo podobno rešitev kot v našem začetnem primeru. Tako kot pri linearni regresiji iz prejšnjega poglavju, začnemo s razredom, ki razvije kriterijsko funkcijo nad vhodnimi podatki.

```
class LogReg:
    def __init__(self, n_inputs, reg=None, reg_strength=0.0):
        self.weights =
          [Value(random.uniform(-1, 1), label=f"w{i}")
           for i in range(n_inputs)]
        self.b = Value(0.0, label="b")
        self.reg = reg
        self.reg_strength = reg_strength

    def linear(self, x):
        return sum(w * xi for w, xi in zip(self.weights, x)) + self.b

    def __call__(self, x):
        return self.linear(x).sigmoid()

    def parameters(self):
        return self.weights + [self.b]

    def loss(self, xs, ys):
        eps = 1e-8
```

```
        losses = []
        for x, y in zip(xs, ys):
            yhat = self(x)
            y_val = Value(float(y))
            term = -(y_val * (yhat + eps).log() + \
              (1 - y_val) * (1 - yhat + eps).log())
            losses.append(term)
        data_loss = sum(losses) / Value(len(xs))

        if self.reg == "l2" and self.reg_strength > 0:
            l2_penalty = self.reg_strength * sum(w * w for w in self.weights)
            return data_loss + l2_penalty
        return data_loss

    def __repr__(self):
        weights_str = ", ".join(f"w{i}={w.data:.3f}" \
          for i, w in enumerate(self.weights))
        return f"LogReg({weights_str}, b={self.b.data:.3f})"
```

Multinomna logistična regresija

Logistična regresija, ki smo jo obravnavali do sedaj, je namenjena klasifikaciji z dvema razredoma. Pogosto pa se srečamo z nalogami, kjer je možnih razredov več. Na primer: želimo napovedati, katero prevozno sredstvo bo posameznik izbral, ali pa v katero kategorijo spada določen dokument. Takšne probleme lahko obravnavamo z razširitvijo logistične regresije, ki ji pravimo *multinomna logistična regresija*. Ta model je poseben primer posplošenih linearnih modelov, kjer ciljna spremenljivka sledi kategorični porazdelitvi.

Naj bo ciljna spremenljivka $y \in \{1, 2, \ldots, m\}$, kjer je m število razredov. Za vsak razred j definiramo linearni napovednik

$$z_j = x^T \theta_j.$$

Ker želimo dobiti verjetnosti, ki so nenegativne in seštejejo v 1, uporabimo funkcijo

$$P(y = j \mid x) = \frac{e^{z_j}}{\sum_{l=1}^{m} e^{z_l}},$$

ki ji pravimo *softmax*. Ta funkcija preslika vektor realnih vrednosti $(z_1, \ldots, z_m)$ v verjetnostni vektor.

Model ni enolično določen, saj lahko vsem z_j prištejemo isto konstanto, ne da bi se verjetnosti spremenile. Zato običajno en razred izberemo kot referenčni razred in njegov linearni napovednik nastavimo na nič:

$$z_m = 0.$$

Model ima tako $(m-1)$ vektorjev parametrov. Multinomna logistična regresija je posplošeni linearni model, kjer ciljna spremenljivka sledi kategorični porazdelitvi, je linearni napovednik $z_j = x^T\theta_j$, je povezovalna funkcija posplošeni logit, katere inverz je funkcija softmax.

Model lahko zapišemo tudi kot

$$y \mid x \sim \text{Categorical}(p_1(x), \ldots, p_m(x)),$$

kjer velja

$$p_j(x) = \frac{e^{x^T\theta_j}}{\sum_{l=1}^{m} e^{x^T\theta_l}}.$$

Če imamo le dva razreda ($m = 2$), se softmax poenostavi v sigmoidno funkcijo in dobimo običajno logistično regresijo.

Parametre modela ocenimo z maksimizacijo logaritemskega verjetja

$$\ell(\theta) = \sum_{i=1}^{n} \log P(y_i \mid x_i),$$

kar vodi do kriterijske funkcije, ki jo poznamo kot večrazredno križno entropijo. Tako kot pri logistični regresiji analitične oblike rešitve ni, zato parametre ocenjujemo numerično, npr. z gradientnim sestopom.

Ordinalna logistična regresija

V nekaterih problemih ciljna spremenljivka ni le diskretna, temveč imajo njene vrednosti tudi naravni vrstni red. Takšne spremenljivke imenujemo *ordinalne*. Primer so odgovori v anketah (npr. "se ne strinjam", "nevtralno", "se strinjam") ali ocene (npr. od 1 do 5). Takšne podatke bi lahko obravnavali z multinomno logistično regresijo, vendar ta ne upošteva vrstnega reda med razredi. Ordinalna logistična regresija ta vrstni red eksplicitno modelira, zato je pogosto bolj učinkovita in interpretabilna.

Osnovna ideja modela je, da obstaja latentna (neopažena) zvezna spremenljivka z, ki je linearno odvisna od značilk:

$$z = x^T\theta.$$

Opazovana ordinalna spremenljivka y je določena glede na to, v kateri interval pade latentna spremenljivka z. Te intervale določajo pragovi (meje) $t_1, t_2, \ldots, t_{m-1}$:

$$y = j \quad \text{če} \quad t_{j-1} < z \leq t_j,$$

kjer za konvencijo vzamemo $t_0 = -\infty$ in $t_m = \infty$.

Da dobimo verjetnosti, predpostavimo, da je latentna spremenljivka obremenjena z logističnim šumom. To vodi do modela, kjer je verjetnost, da je y manjši ali enak določenemu razredu, enaka

$$P(y \leq j \mid x) = \frac{1}{1 + e^{-(t_j - x^T\theta)}}.$$

Ta izraz predstavlja kumulativno verjetnost, zato model pogosto imenujemo tudi *kumulativni logit model*. Verjetnosti posameznih razredov dobimo kot razlike med zaporednimi kumulativnimi verjetnostmi:

$$P(y = j \mid x) = P(y \leq j \mid x) - P(y \leq j - 1 \mid x).$$

Ordinalna logistična regresija je posplošeni linearni model, kjer ciljna spremenljivka sledi kategorični porazdelitvi z urejenimi razredi, je linearni napovednik enak $x^T\theta$, in je povezovalna funkcija je kumulativni logit.

Model uporablja en sam vektor parametrov θ, pragovi t_j pa določajo meje med razredi. Zaradi tega ima model manj parametrov kot multinomna logistična regresija.

Parametre modela ocenimo z maksimizacijo logaritemskega verjetja

$$\ell(\theta) = \sum_{i=1}^{n} \log P(y_i \mid x_i),$$

kar ponovno vodi do optimizacijskega problema brez zaprte oblike rešitve, zato parametre ocenjujemo numerično.