Classification

Hui Chen

MIT Sloan

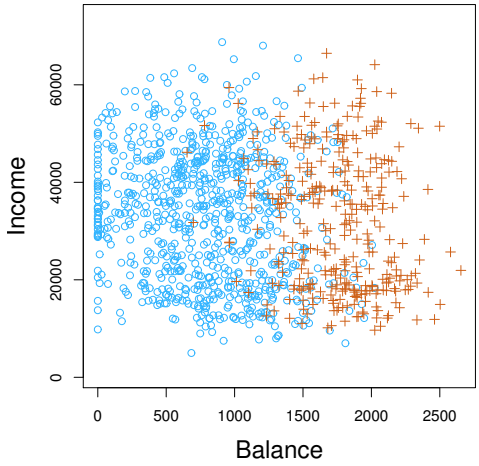
15.450, Spring 2023

Outline

1. Classification
2. Logistic Regression
3. K-Nearest Neighbors

Example: Credit Card Defaults

Source: ISL



Classification

- Predicting binary or categorial data is referred to as classification.
- Suppose y takes the value of 1 or 0 (e.g., *default* or *solvent*). We want to predict y using p features, $\mathbf{x}$.
- Recall the multiple regression model:

$$y_i = \beta' \mathbf{x}_i + \varepsilon$$

  ↪ Not a good model, because the predicted value $\hat{y}_i$ could be outside of $[0, 1]$.

- Instead, we can model the (conditional) probability of y taking on 0 or 1, given the information contained in vector $\mathbf{x}$.

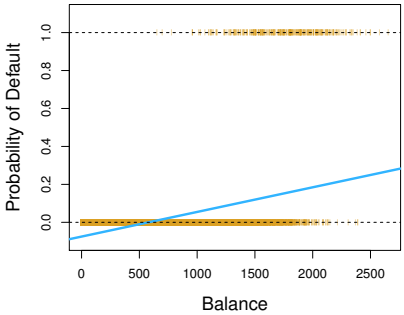
$$p(y = 1|\mathbf{x}) = F(\mathbf{x}, \theta)$$
$$p(y = 0|\mathbf{x}) = 1 - F(\mathbf{x}, \theta)$$

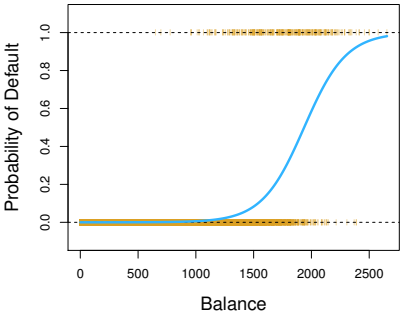
- Notice that we are still effectively modeling the expected value of y:

$$\mathrm{E}[y|\mathbf{x}] = p(y = 1|\mathbf{x}) = F(\mathbf{x}, \theta)$$

Example: Credit Card Defaults

Source: ISL





Classification

- Many potential choices for $F()$. Would like F to be between 0 and 1 to give proper probabilities.
- Popular examples:
  - ↪ Probit:

$$F(\mathbf{x}, \theta) = \Phi(\theta' \mathbf{x})$$

  - ↪ Logit:

$$F(\mathbf{x}, \theta) = sigm(\theta' \mathbf{x}) = \frac{e^{\theta' \mathbf{x}}}{1 + e^{\theta' \mathbf{x}}} = \frac{1}{1 + e^{-\theta' \mathbf{x}}}$$

Outline

1. Classification
2. Logistic Regression
3. K-Nearest Neighbors

- **Log odds ratio**:

$$LO \equiv \log \frac{p(y=1|\mathbf{x})}{p(y=0|\mathbf{x})}$$

- LO is continuous and ranges between $-\infty$ and $+\infty$.
- We can directly model LO as linear function of $\mathbf{x}$:

$$LO = \theta'\mathbf{x}$$

- This implies the logit model:

$$p(y=1|\mathbf{x}) = \frac{e^{\theta'\mathbf{x}}}{1+e^{\theta'\mathbf{x}}}$$

Q: Where is the error term?

Logistic Regression: Estimation with MLE

- The observations $y_1, y_2, \cdots, y_n$ are assumed to be IID binomial conditional on the features.
- Likelihood function for y_i:

$$p(y_i|\mathbf{x}_i, \theta) = F(\mathbf{x}_i, \theta)^{y_i}(1 - F(\mathbf{x}_i, \theta))^{1-y_i} = \left(\frac{e^{\theta'\mathbf{x}_i}}{1 + e^{\theta'\mathbf{x}_i}}\right)^{y_i}\left(\frac{1}{1 + e^{\theta'\mathbf{x}_i}}\right)^{1-y_i}$$

- Log-likelihood for Y:

$$\begin{aligned}\mathscr{L}(\theta) &= \log L(y_1, \cdots, y_n|\mathbf{x}_1, \cdots, \mathbf{x}_n, \theta) = \sum_{i=1}^{n}\left(y_i \log F(\mathbf{x}_i, \theta) + (1 - y_i)\log(1 - F(\mathbf{x}_i, \theta))\right) \\ &= \sum_{i=1}^{n}\left(y_i \log \frac{F(\mathbf{x}_i, \theta)}{1 - F(\mathbf{x}_i, \theta)} + \log(1 - F(\mathbf{x}_i, \theta))\right) = \sum_{i=1}^{n}\left(y_i\theta'\mathbf{x}_i - \log\left(1 + e^{\theta'\mathbf{x}_i}\right)\right)\end{aligned}$$

- No closed-form solution in general, but "easy" to solve numerically, e.g., using *glm* or the *glmnet* package in R.

Classification and Confusion Matrix

- After estimating the logistic regression coefficients $\widehat{\theta}$, we can make predictions:

$$\widehat{p}(y = 1|\mathbf{x}) = \frac{e^{\widehat{\theta}'\mathbf{x}}}{1 + e^{\widehat{\theta}'\mathbf{x}}}$$

- We can also classify the data based on a cutoff rule with threshold $\bar{p}$:

$$\text{Predict } y = 1 \text{ if } \widehat{p}(y = 1|\mathbf{x}) > \bar{p}; \text{ otherwise } y = 0.$$

- How good are our predictions?
  - ↪ Use the confusion matrix to compare the predicted and true classes of the observations.

Confusion Matrix

		Predicted class		
		−	+	Total
True class	−	True Neg. (TN)	False Pos. (FP)	N
	+	False Neg. (FN)	True Pos. (TP)	P
	Total	N^*	P^*	

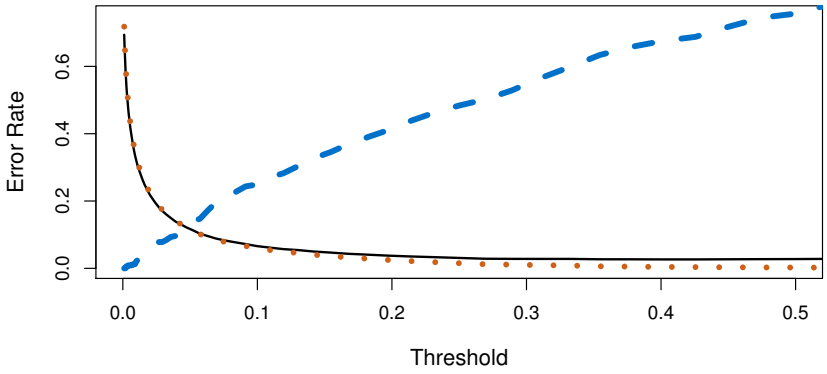
- Type I error rate (false positive) = $\frac{FP}{N}$
- Type II error rate (false negative) = $\frac{FN}{P}$

Confusion matrix for predicting defaults

		Predicted default		
		No	Yes	Total
True default status	No	9,644	23	9,667
	Yes	252	81	333
	Total	9,896	104	10,000

- Type I error rate (false positive):
- Type II error rate (false negative):

Threshold $\bar{p}$ and Type I/II errors



Economic loss function

- Suppose the dollar costs of making Type-I and Type II errors are L_I and L_{II}.
  - ↪ What determines the magnitudes of L_I and L_{II}?
- Economic objective: Maximize the expected profit per application.

$$\underbrace{\frac{N}{N+P}\left(1-\frac{FP}{N}\right)L_I}_{\text{profit of lending to a good borrower}} + \underbrace{\frac{P}{N+P}\frac{FN}{P}(-L_{II})}_{\text{loss of lending to a bad borrower}}$$

$$= \underbrace{\frac{N}{N+P}L_I}_{\text{maximum profit}} - \underbrace{\left(\frac{N}{N+P}\frac{FP}{N}L_I + \frac{P}{N+P}\frac{FN}{P}L_{II}\right)}_{\text{expected loss}}$$

- Optimal decision rule: Choose $\overline{p}$ to minimize the expected economic losses.

Multi-class Logistic Regression

- Suppose y takes value ℓ, where $\ell \in \{1, \cdots, C\}$.
- Model the log odds ratio for class c relative to class C as:

$$\log \frac{p(y=c|\mathbf{x})}{p(y=C|\mathbf{x})} = \theta_c' \mathbf{x}$$

  ↪ Normalized against the C-th class; this choice does not matter.
  ↪ Notice that θ_c is a $p \times 1$ vector that changes with c.

- This implies

$$p(y=c|\mathbf{x},\theta) = \frac{\exp(\theta_c'\mathbf{x})}{1+\sum_{\ell=1}^{C-1}\exp(\theta_\ell'\mathbf{x})}, \quad c = 1, \cdots, C-1$$

$$p(y=C|\mathbf{x},\theta) = \frac{1}{1+\sum_{\ell=1}^{C-1}\exp(\theta_\ell'\mathbf{x})}$$

Example: Predicting Corporate Defaults

- At firm level: the events of default, survival, exit, merger, etc. are categorical data: $Y_{it} = \{0, 1\}$
- Pre-default, we observe firm and market information: X_{it}
- Forecasting model: Connect firm and market characteristics to the likelihoods of default.
  - ↪ Logit and Probit
  - ↪ Rating-migration model (e.g. S&P)
  - ↪ Hazard model

- Marginal probability of default over the next period:

$$P_{t-1}(Y_{it} = 1) = \frac{1}{1 + \exp(-\beta' X_{i,t-1})}$$

  ↪ $X_{i,t-1}$: vector of explanatory variables (covariates)

- Long horizon:

$$P_{t-1}(Y_{i,t-1+j} = 1 \,|\, Y_{i,t-2+j} = 0) = \frac{1}{1 + \exp(-\beta_j' X_{i,t-1})}$$

  ↪ How to calculate $P_{t-1}(Y_{i,t} = 0, \cdots, Y_{i,t+j} = 0)$?
  ↪ Cumulative default probability no longer has logit form.
  ↪ Not fully utilizing the information in the dynamics of $X_{i,t}$.

Example: Campbell, Hilscher, Szilagyi (JF 2008)

Table I

Number of Bankruptcies and Failures per Year

This table lists the total number of active firms, bankruptcies, and failures for every year of our sample period. The number of active firms is computed by averaging over the numbers of active firms across all months of the year.

Year	Active Firms	Bankruptcies	(%)	Failures	(%)
1963	1,281	0	0.00	0	0.00
1964	1,357	2	0.15	2	0.15
1965	1,436	2	0.14	2	0.14
1966	1,513	1	0.07	1	0.07
1967	1,598	0	0.00	0	0.00
1968	1,723	0	0.00	0	0.00
1969	1,885	0	0.00	0	0.00
1970	2,067	5	0.24	5	0.24
1971	2,199	4	0.18	4	0.18
1972	2,650	8	0.30	8	0.30
1973	3,964	6	0.15	6	0.15
1974	4,002	18	0.45	18	0.45
1975	4,038	5	0.12	5	0.12
1976	4,101	14	0.34	14	0.34
1977	4,157	12	0.29	12	0.29
1978	4,183	14	0.33	15	0.36
1979	4,222	14	0.33	14	0.33
1980	4,342	26	0.60	26	0.60
1981	4,743	23	0.48	23	0.48
1982	4,995	29	0.58	29	0.58
1983	5,380	50	0.93	50	0.93
1984	5,801	73	1.26	74	1.28
1985	5,912	76	1.29	77	1.30
1986	6,208	95	1.53	95	1.53
1987	6,615	54	0.82	54	0.82
1988	6,686	84	1.26	85	1.27
1989	6,603	74	1.12	78	1.18
1990	6,515	80	1.23	82	1.26

Example: Campbell, Hilscher, Szilagyi (JF 2008)

Table III

Logit Regressions of Bankruptcy/Failure Indicator on Predictor Variables

This table reports results from logit regressions of the bankruptcy and failure indicators on predictor variables. The data are constructed such that all of the predictor variables are observable at the beginning of the month over which bankruptcy or failure is measured. The absolute value of *z*-statistics is reported in parentheses. *denotes significant at 5%, **denotes significant at 1%.

	Model 1			Model 2		
Dependent variable:	Bankruptcy	Failure	Failure	Bankruptcy	Failure	Failure
Sample period:	1963–1998	1963–1998	1963–2003	1963–1998	1963–1998	1963–2003
NITA	−14.05	−13.79	−12.78			
	(16.03)**	(17.06)**	(21.26)**			
NIMTAAVG				−32.52	−32.46	−29.67
				(17.65)**	(19.01)**	(23.37)**
TLTA	5.38	4.62	3.74			
	(25.91)**	(26.28)**	(32.32)**			
TLMTA				4.32	3.87	3.36
				(22.82)**	(23.39)**	(27.80)**
EXRET	−3.30	−2.90	−2.32			
	(12.12)**	(11.81)**	(13.57)**			
EXRETAVG				−9.51	−8.82	−7.35
				(12.05)**	(12.08)**	(14.03)**
SIGMA	2.15	2.28	2.76	0.920	1.15	1.48
	(16.40)**	(18.34)**	(26.63)**	(6.66)**	(8.79)**	(13.54)**
RSIZE	−0.188	−0.253	−0.374	0.246	0.169	0.082
	(5.56)**	(7.60)**	(13.26)**	(6.18)**	(4.32)**	(2.62)**
CASHMTA				−4.89	−3.22	−2.40
				(7.96)**	(6.59)**	(8.64)**
MB				0.099	0.095	0.054
				(6.72)**	(6.76)**	(4.87)**
PRICE				−0.882	−0.807	−0.937
				(10.39)**	(10.09)**	(14.77)**
Constant	−15.21	−15.41	−16.58	−7.65	−8.45	−9.08
	(39.45)**	(40.87)**	(50.92)**	(13.66)**	(15.63)**	(20.84)**
Observations	1,282,853	1,302,564	1,695,036	1,282,853	1,302,564	1,695,036
Failures	797	911	1,614	797	911	1,614
Pseudo-R^2	0.260	0.258	0.270	0.299	0.296	0.312

Example: Campbell, Hilscher, Szilagyi (JF 2008)

Table IV

Logit Regressions of Failure Indicator on Lagged Variables

This table takes our best-model variables (model 2 in Table III) and reports their predictive power for lags of 6, 12, 24, and 36 months. The dependent variable is failure and the sample period is 1963 to 2003. The absolute value of z-statistics is reported in parentheses. *denotes significant at 5%, **denotes significant at 1%.

Lag (Months)	0	6	12	24	36
NIMTAAVG	−29.67	−23.92	−20.26	−13.23	−14.06
	(23.37)**	(21.82)**	(18.09)**	(10.50)**	(9.77)**
TLMTA	3.36	2.06	1.42	0.917	0.643
	(27.80)**	(22.63)**	(16.23)**	(9.85)**	(6.25)**
EXRETAVG	−7.35	−7.79	−7.13	−5.61	−2.56
	(14.03)**	(15.97)**	(14.15)**	(10.14)**	(4.14)**
SIGMA	1.48	1.27	1.41	1.52	1.33
	(13.54)**	(14.57)**	(16.49)**	(16.92)**	(13.54)**
RSIZE	0.082	0.047	−0.045	−0.132	−0.180
	(2.62)**	(2.02)*	(2.09)*	(6.19)**	(8.03)**
CASHMTA	−2.40	−2.40	−2.13	−1.37	−1.41
	(8.64)**	(9.77)**	(8.53)**	(5.09)**	(4.61)**
MB	0.054	0.047	0.075	0.108	0.125
	(4.87)**	(4.22)**	(6.33)**	(7.92)**	(8.17)**
PRICE	−0.937	−0.468	−0.058	0.212	0.279
	(14.77)**	(10.36)**	(1.40)	(4.96)**	(6.00)**
Constant	−9.08	−8.07	−9.16	−10.23	−10.53
	(20.84)**	(25.00)**	(30.89)**	(34.48)**	(33.53)**
Observations	1,695,036	1,642,006	1,565,634	1,384,951	1,208,610
Failures	1,614	2,008	1,968	1,730	1,467
Pseudo-R^2	0.312	0.188	0.114	0.061	0.044

Example: Rating-Migration Model

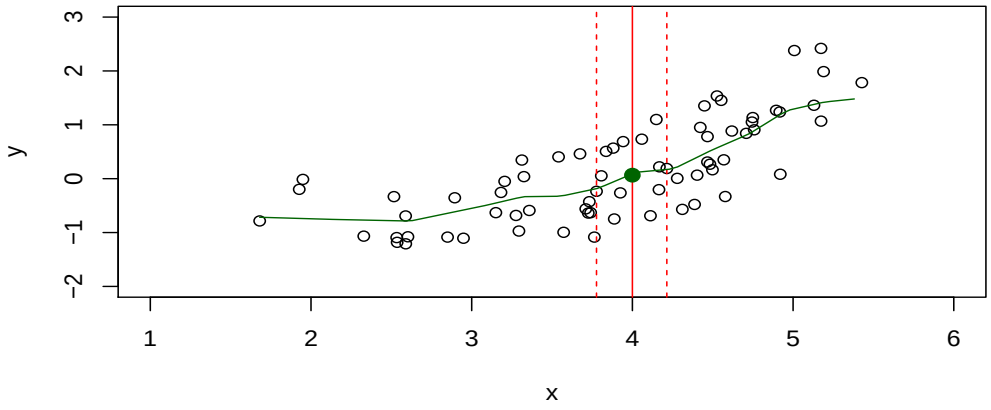
- How to forecast the probability of default of a particular bond/firm?
- Credit rating agencies provide credit rating to bonds, which provide (their) assessment of their probability of default
- Example: Standard and Poor's rating transition matrix

Initial Rating	Rating at the end of the year							
	AAA	AA	A	BBB	BB	B	CCC	D
AAA	0.8910	0.0963	0.0078	0.0019	0.0030	0.0000	0.0000	0.0000
AA	0.0086	0.9010	0.0747	0.0099	0.0029	0.0029	0.0000	0.0000
A	0.0009	0.0291	0.8894	0.0649	0.0101	0.0045	0.0000	0.0009
BBB	0.0006	0.0043	0.0656	0.8427	0.0644	0.0160	0.0018	0.0045
BB	0.0004	0.0022	0.0079	0.0719	0.7764	0.1043	0.0127	0.0241
B	0.0000	0.0019	0.0031	0.0066	0.0517	0.8246	0.0435	0.0685
CCC	0.0000	0.0000	0.0116	0.0116	0.0203	0.0754	0.6493	0.2319
D	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000

Outline

1. Classification
2. Logistic Regression
3. K-Nearest Neighbors

A non-parametric approach



- Instead of fitting a (linear) function to the data, try predicting the value for y based on the observations nearby.

K-Nearest Neighbors

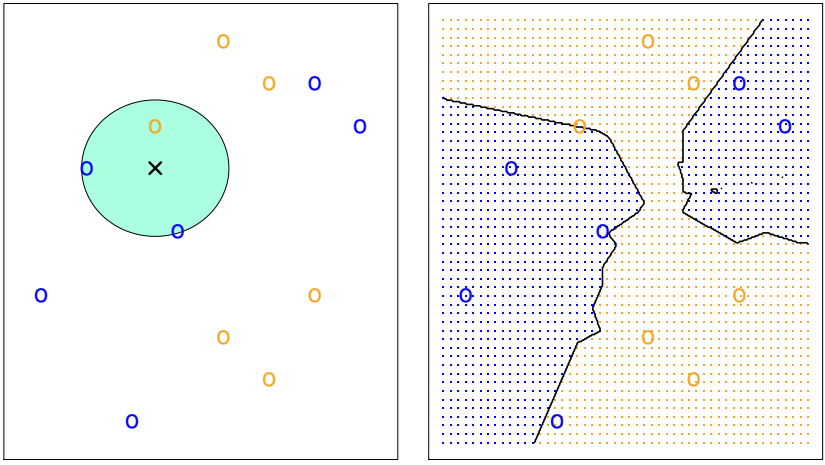
- One way to predict whether a borrower with attributes $\mathbf{x}_0$ will default or not is to look at how often the other borrowers who are "similar" to him (neighbors) default in the data.
- KNN applies the following procedure:
  1. Identify K points in the training data that are closest to $\mathbf{x}_0$, denoted by $\mathcal{N}_0$.
  2. The conditional probability of y belonging to class c is:

$$p(y = c|\mathbf{x} = \mathbf{x}_0) = \frac{1}{K} \sum_{i \in \mathcal{N}_0} I(y_i = c)$$

  3. Pick a set of cutoff probabilities, $\overline{p}^c$, $c = 1, \cdots, C-1$, such that we classify y as in class c if $p(y = c|\mathbf{x} = \mathbf{x}_0) > \overline{p}^c$.
  4. Obtain the KNN decision boundary in the feature space.

K-Nearest Neighbors

Source: ISL



How to measure the distance?

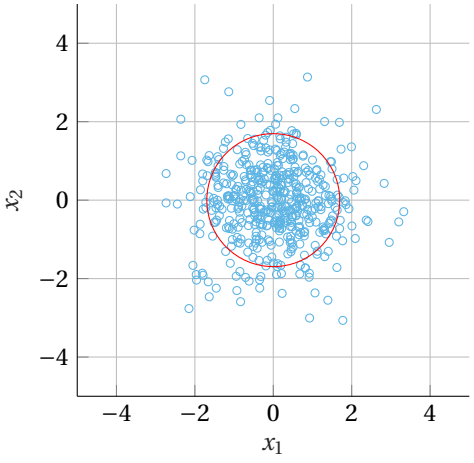
- How to measure the distance between data points?
- Real-valued features: First standardize each of the features to have zero mean and unit variance. Then use the Euclidean (squared) distance in feature space:

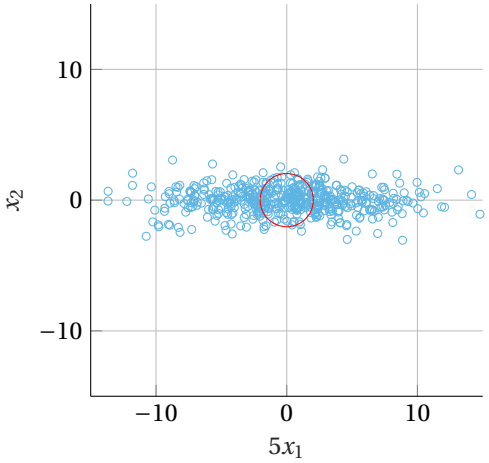
$$d_i = \|\mathbf{x}_i - \mathbf{x}_0\|$$

- Qualitative features:

$$d_i = \begin{cases} 0 & \mathbf{x}_i = \mathbf{x}_0 \\ 1 & \mathbf{x}_i \neq \mathbf{x}_0 \end{cases}$$

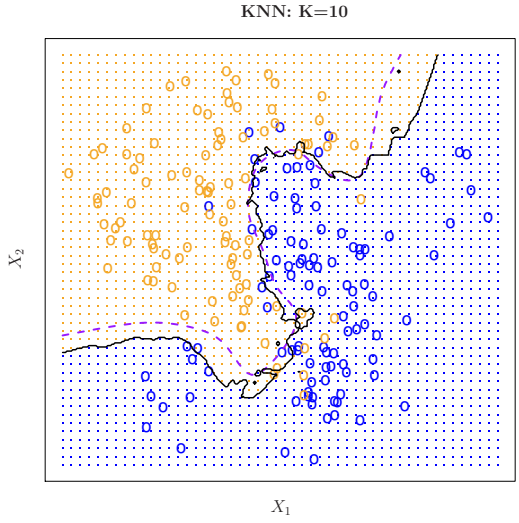
Size matters





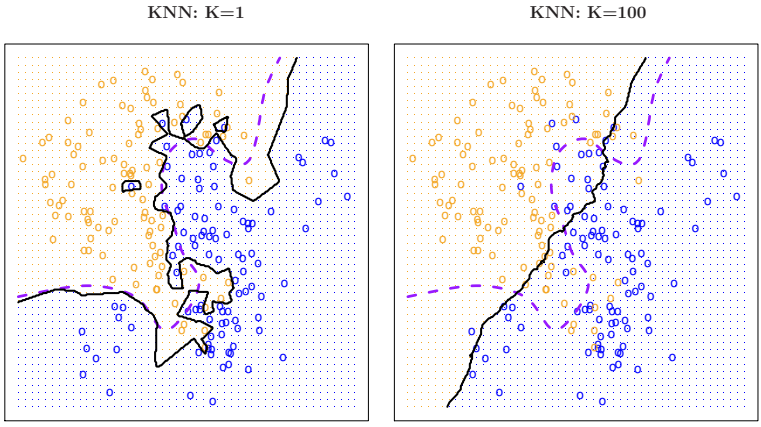
K-Nearest Neighbors

Source: ISL



The choice for K

Source: ISL



- How to choose K? Cross validation (later).

Default Forecasting in FinTech

- Big data and machine learning methods
  - ↪ No reason for the relation between log default prob and covariates to be linear.
  - ↪ Superior ability to predict defaults has become the core competitive advantage for some FinTech companies (e.g., Alibaba, LendingClub ...)
- Example: Alibaba's microlending platform
  - ↪ Borrowers: over 400k/per year
  - ↪ Total credit: exceeding $40 bil
  - ↪ Loan size: as small as 100 RMB ($15)
  - ↪ 100% unsecured
  - ↪ NPL: < 1%
  - ↪ Beyond standard info: user ratings, "punishment records", repeat buyers ...

- The frontier:
  - ↪ Correlate defaults: Jointly model many borrowers.
  - ↪ "Small data": Large XS helps little with estimating aggregate default risk.
  - ↪ Robust credit screening

Summary

- Logistic regression
- Confusion matrix, Type I/II errors
- KNN
- ISL Chapter 2.2.3, 4.1-4.3