Computational Pattern Analysis and Statistical Learning

Lecture 5: Supervised learning

Tijl De Bie, Konstantin Tretyakov
(Largely based on joint work with Nello Cristianini and John Shawe-Taylor)

Tartu, Estonia

November 2006

1. Lecture 5A: Regression and classification
   - Linear regression
   - Fisher's discriminant analysis
   - Support Vector Machines
2. Lecture 5B: Kernel regression and classification, and stability analysis
   - Kernel ridge regression
   - How to 'kernelise' an algorithm? – you should know now
   - Kernel support vector machines
   - Statistical analysis of ridge regression
3. Wrap-up Lecture 5

Overview

- Recapitulation of ridge regression – now with offset
- Fisher's discriminant analysis
- Support Vector Machines

Least squares regression

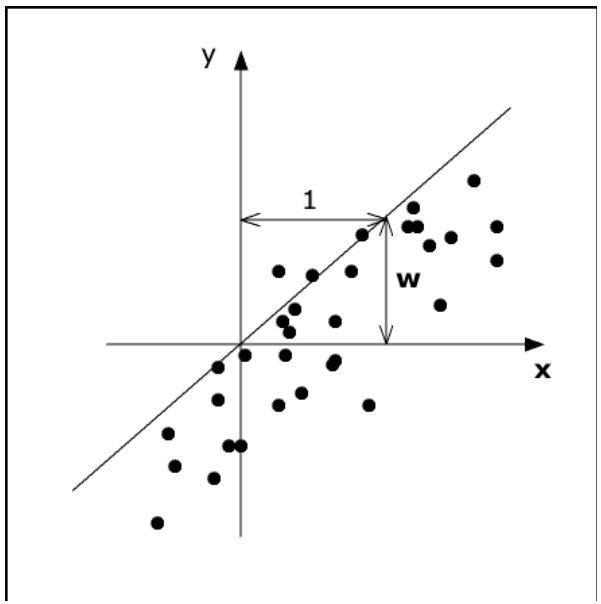
- We want to approximate y_i as a linear function of $\mathbf{x}_i$
- In terms of a weight vector $\mathbf{w}$, this means: $y_i \approx \mathbf{x}_i'\mathbf{w}$, or, $\|y_i - \mathbf{x}_i'\mathbf{w}\| \approx 0$
- Pattern function is parameterised by $\mathbf{w}$ (note the $-$ sign):

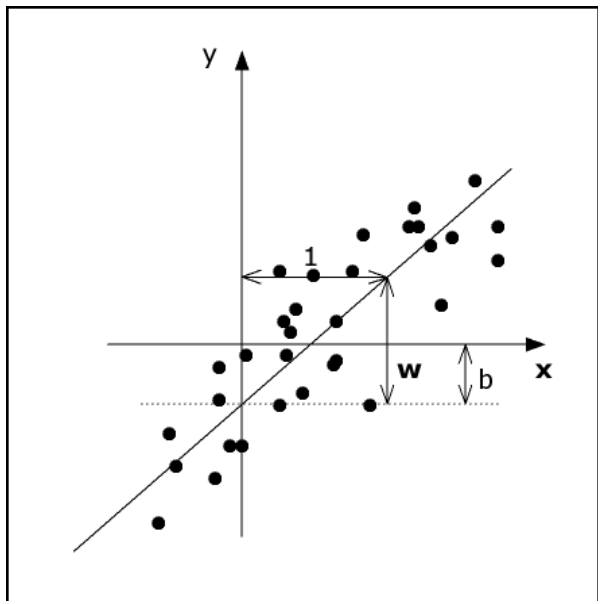
$$\pi_{\mathbf{w}}(Z) = -\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \mathbf{x}_i'\mathbf{w}\right)^2 = -\frac{1}{n}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2$$

- Formal pattern recognition problem:

$$\max_{\mathbf{w}} \pi_{\mathbf{w}}(Z) \Leftrightarrow \max_{\mathbf{w}} -\frac{1}{n}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2 \Leftrightarrow \min_{\mathbf{w}} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2$$

Least squares regression with offset





Least squares regression with offset

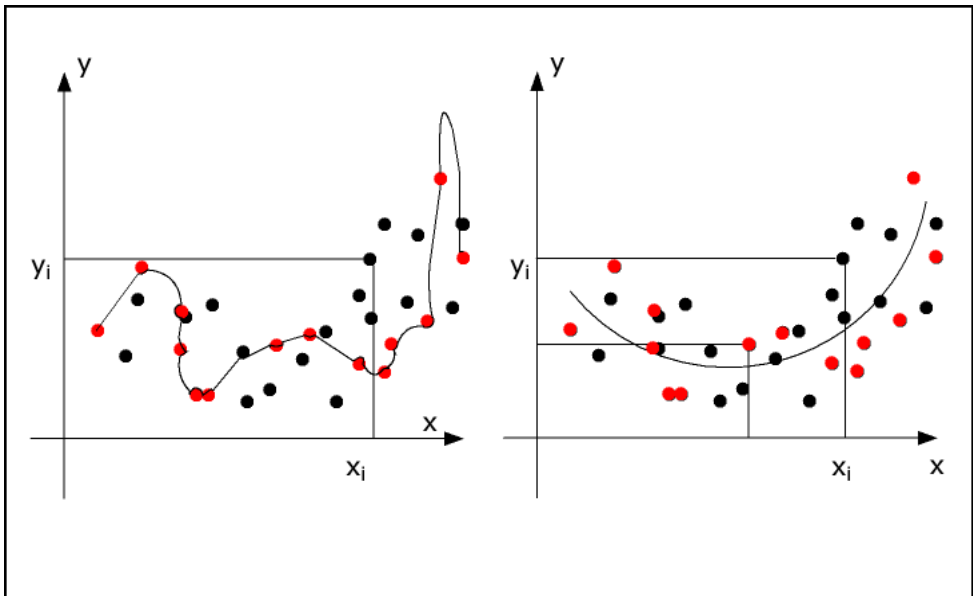
- We want to approximate y_i as an *affine* function of $\mathbf{x}_i$
- In terms of a weight vector $\mathbf{w}$ and offset b, this means: $y_i \approx \mathbf{x}_i'\mathbf{w} + b$, or, $\|y_i - (\mathbf{x}_i'\mathbf{w} + b)\| \approx 0$
- Pattern function is parameterised by $\mathbf{w}$ (note the $-$ sign):

$$\pi_{\mathbf{w},b}(Z) = -\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \left(\mathbf{x}_i'\mathbf{w} + b\right)\right)^2 = -\frac{1}{n}\|\mathbf{y} - \mathbf{X}\mathbf{w} - \mathbf{1}b\|^2$$

- Formal pattern recognition problem:

$$\max_{\mathbf{w},b} \pi_{\mathbf{w},b}(Z) \Leftrightarrow \max_{\mathbf{w},b} -\frac{1}{n}\|\mathbf{y} - \mathbf{X}\mathbf{w} - \mathbf{1}b\|^2 \Leftrightarrow \min_{\mathbf{w}} \|\mathbf{y} - \mathbf{X}\mathbf{w} - \mathbf{1}b\|^2$$

Ridge regression with offset



- Danger for overfitting (usually not in 1/low-dimensional regression, but in high-dimensional spaces such as when using kernel trick to do nonlinear regression)
- Capacity control: regularise by additionally controlling $C\left(\pi_{\mathbf{w},b}\right) = \|\mathbf{w}\|^2$

Ridge regression with offset

$$\min_{\mathbf{w}} \|\mathbf{y} - \mathbf{X}\mathbf{w} - \mathbf{1}b\|^2 + \gamma \|\mathbf{w}\|^2$$

- Solve by taking gradient w.r.t. $\mathbf{w}$, and derivative w.r.t b, and equating to 0:

$$\begin{cases} (\gamma \mathbf{I} + \mathbf{X}'\mathbf{X})\,\mathbf{w} + \mathbf{X}'\mathbf{1}b - \mathbf{X}'\mathbf{y} &= \mathbf{0} \\ \mathbf{1}'\mathbf{X}\mathbf{w} + \mathbf{1}'\mathbf{1}b - \mathbf{1}'\mathbf{y} &= 0 \end{cases}$$

- Solved by a linear system of equations:

$$\begin{pmatrix} (\gamma \mathbf{I} + \mathbf{X}'\mathbf{X}) & \mathbf{X}'\mathbf{1} \\ \mathbf{1}'\mathbf{X} & \mathbf{1}'\mathbf{1} \end{pmatrix} \begin{pmatrix} \mathbf{w} \\ b \end{pmatrix} = \begin{pmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{1}'\mathbf{y} \end{pmatrix}$$

Fisher's discriminant analysis

- Let's assume *binary* classification: $y_i \in \{-1, 1\}$
- Pattern function: learn classifier as thresholded linear function? $y \leftarrow \text{sign}(\mathbf{x}'\mathbf{w} + b)$
- Then:

$$-g^*_{\pi_{\mathbf{w},b}}(\mathbf{x}, y) = \left(\frac{1 - \text{sign}\left(y\left(\mathbf{x}'\mathbf{w} + b\right)\right)}{2}\right)^2$$

- However, this is hard to optimise... non-convex!
- Hence, use a convex upper bound:

$$-g_{\pi_{\mathbf{w},b}}(\mathbf{x}, y) = \left(1 - y\left(\mathbf{x}'\mathbf{w} + b\right)\right)^2$$

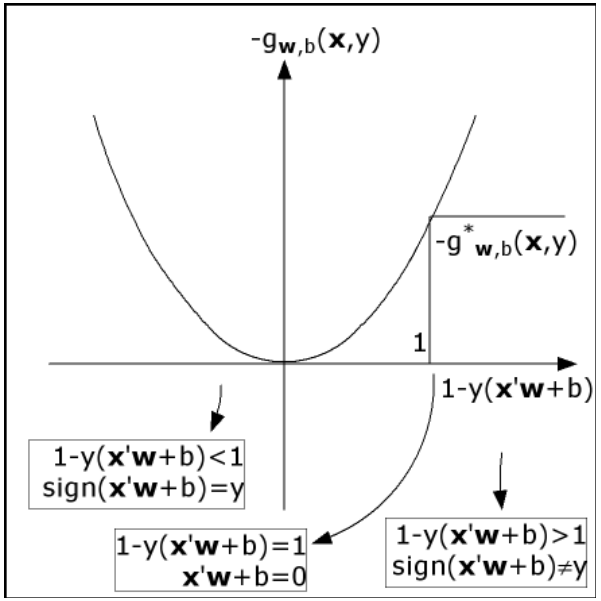
Fisher's discriminant analysis

- Ideal:

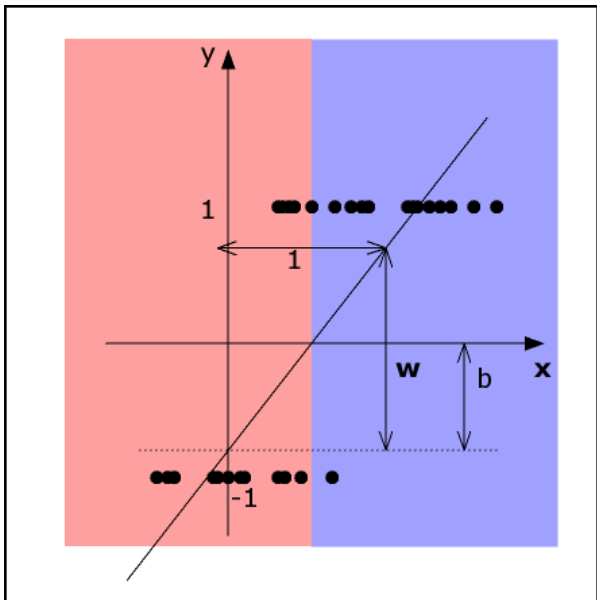
$$-g^*_{\pi_{\mathbf{w},b}}(\mathbf{x}, y) = \left(\frac{1 - \text{sign}\left(y\left(\mathbf{x}'\mathbf{w} + b\right)\right)}{2}\right)^2$$

- Convex upper bound:

$$-g_{\pi_{\mathbf{w},b}}(\mathbf{x}, y) = \left(1 - y\left(\mathbf{x}'\mathbf{w} + b\right)\right)^2$$

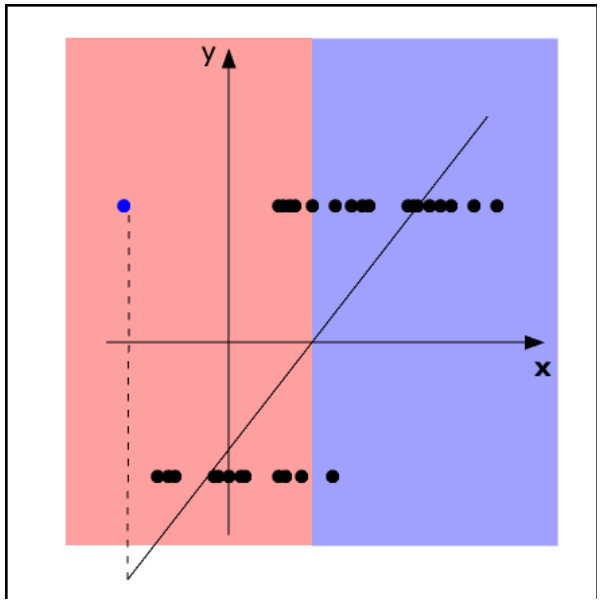


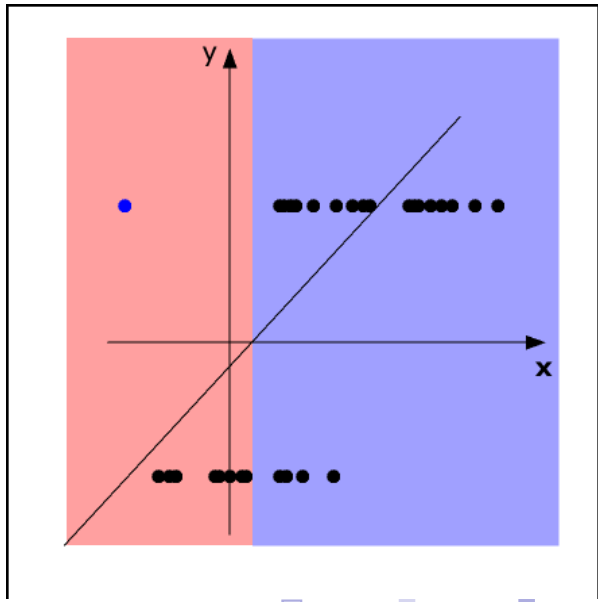
- Note, for y binary,
  $-g_{\pi_{\mathbf{w},b}}(\mathbf{x}, y) = (1 - y(\mathbf{x}'\mathbf{w} + b))^2 = (y - (\mathbf{x}'\mathbf{w} + b))^2$
- Same as for ridge regression!
- Hence, exact same methodology as for (ridge) regression can be used



Fisher's discriminant analysis

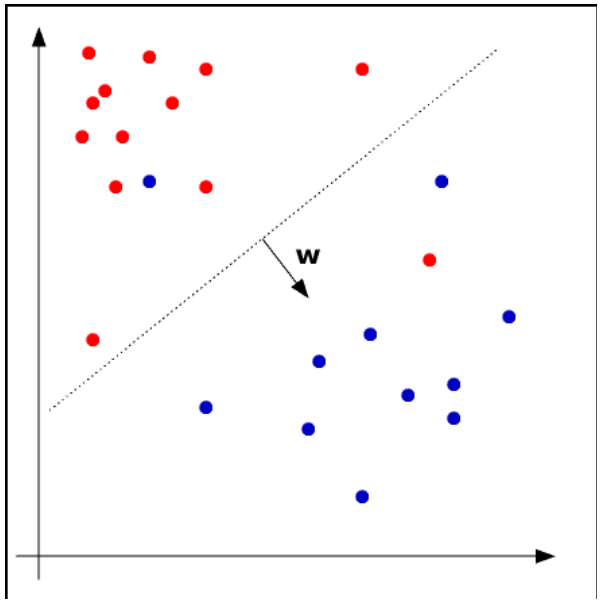
- $\pi_{\mathbf{w},b}(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n} g_{\mathbf{w},b}(\mathbf{x}_i)$ with
  $g_{\mathbf{w},b}(\mathbf{x}_i) = -(y_i - (\mathbf{x}_i'\mathbf{w} + b))^2 = -(1 - y_i(\mathbf{x}_i'\mathbf{w} + b))^2$
- $-g_{\mathbf{w},b}(\mathbf{x}_i)$ is the *cost* associated to each $(\mathbf{x}_i, y_i)$
- Quite sensitive to outliers (quadratic!)

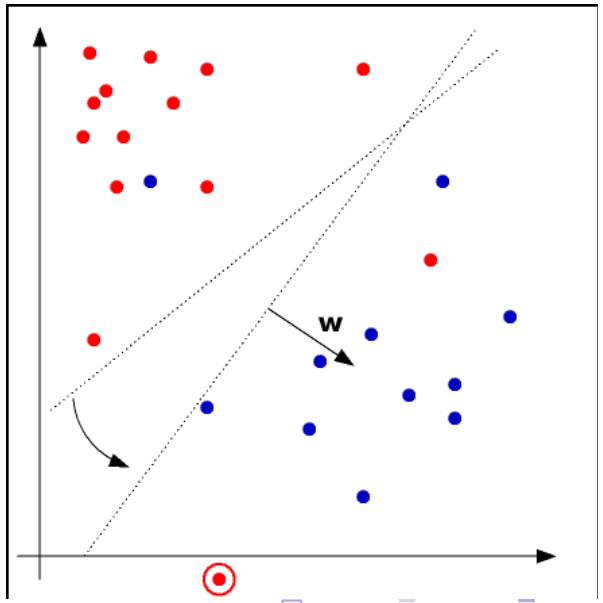




Fisher's discriminant analysis

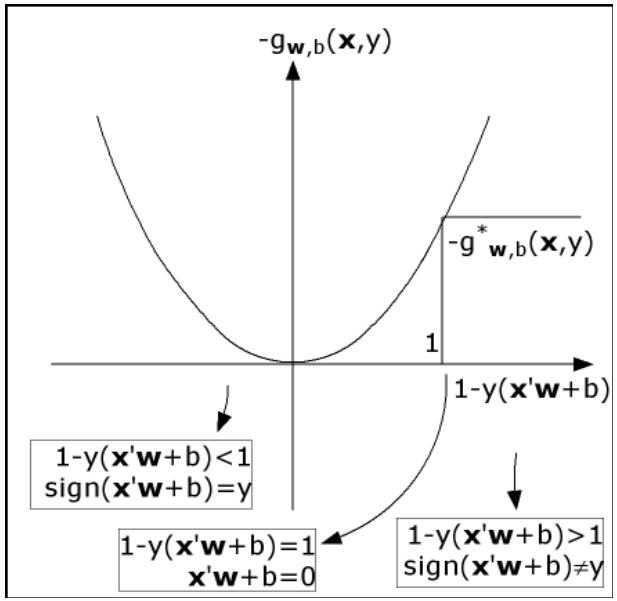
- $\pi_{\mathbf{w},b}(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n} g_{\mathbf{w},b}(\mathbf{x}_i)$ with
  $g_{\mathbf{w},b}(\mathbf{x}_i) = -(y_i - (\mathbf{x}_i'\mathbf{w} + b))^2 = -(1 - y_i(\mathbf{x}_i'\mathbf{w} + b))^2$
- This is the *cost* associated to each $(\mathbf{x}_i, y_i)$
- Quite sensitive to outliers (quadratic!)

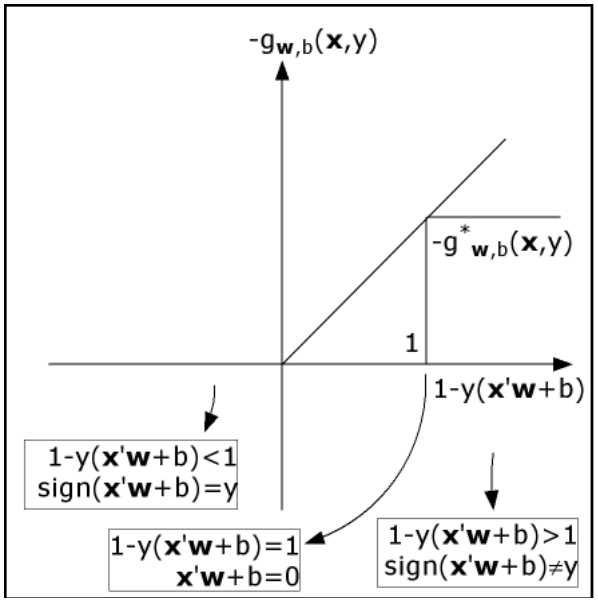




Support Vector Machines for robust regression

- Solution: use another cost (not quadratic), also an upper bound on $-g^*_{\pi_{\mathbf{w},b}}(\mathbf{x}, y) = \left(\frac{1-\text{sign}(y(\mathbf{x}'\mathbf{w}+b))}{2}\right)^2$
- But keep it convex...





Support vector machines

- Averaging pattern function with:

$$g_{\mathbf{w},b}\left(\mathbf{x}_i\right) = -\max\left(0, 1 - y_i\left(\mathbf{x}_i'\mathbf{w} + b\right)\right)$$

- Pattern function itself:

$$\pi_{\mathbf{w},b}\left(\mathbf{X}\right) = -\frac{1}{n}\sum_{i=1}^{n}\max\left(0, 1 - y_i\left(\mathbf{x}_i'\mathbf{w} + b\right)\right)$$

- Capacity functional:

$$C\left(\pi_{\mathbf{w},b}\left(\mathbf{X}\right)\right) = \left\|\mathbf{w}\right\|^2$$

- Pattern recognition problem:

$$\min_{\mathbf{w},b}\frac{1}{n}\sum_{i=1}^{n}\max\left(0, 1 - y_i\left(\mathbf{x}_i'\mathbf{w} + b\right)\right) + \gamma\left\|\mathbf{w}\right\|^2$$

Support vector machines

- Introduce new variables: $\xi_i \geq 0$ and $\xi_i \geq 1 - y_i \left(\mathbf{x}_i' \mathbf{w} + b\right)$
- Then, $\sum_{i=1}^{n} \max\left(0, 1 - y_i \left(\mathbf{x}_i' \mathbf{w} + b\right)\right) = \min_{\boldsymbol{\xi}} \sum \xi_i$
- Hence:

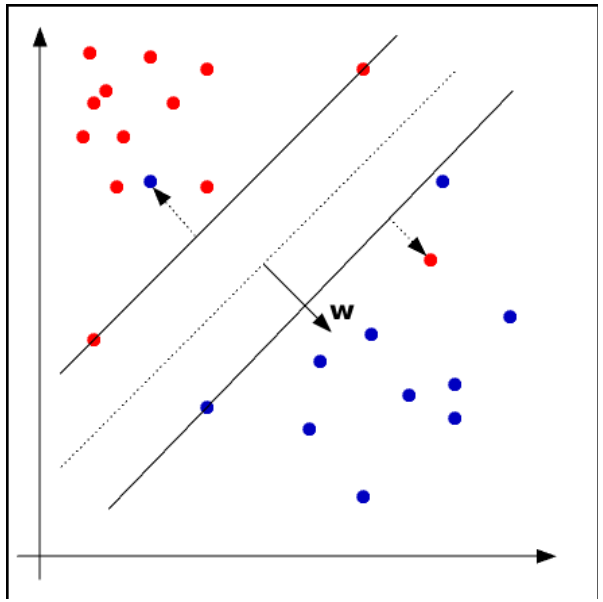
$$\begin{aligned} \min_{\mathbf{w}, b, \boldsymbol{\xi}} \quad & \frac{1}{n} \sum_{i=1}^{n} \xi_i + \gamma \left\|\mathbf{w}\right\|^2 \\ \text{s.t.} \quad & \xi_i \geq 0 \\ & \xi_i \geq 1 - y_i \left(\mathbf{x}_i' \mathbf{w} + b\right) \end{aligned}$$

- This is easy to solve using any quadratic programming toolbox...

Support vector machines

- Property: many $\xi_i = 0$, corresponding to $y_i \left(\mathbf{x}_i'\mathbf{w} + b\right) \geq 1 \Leftrightarrow$

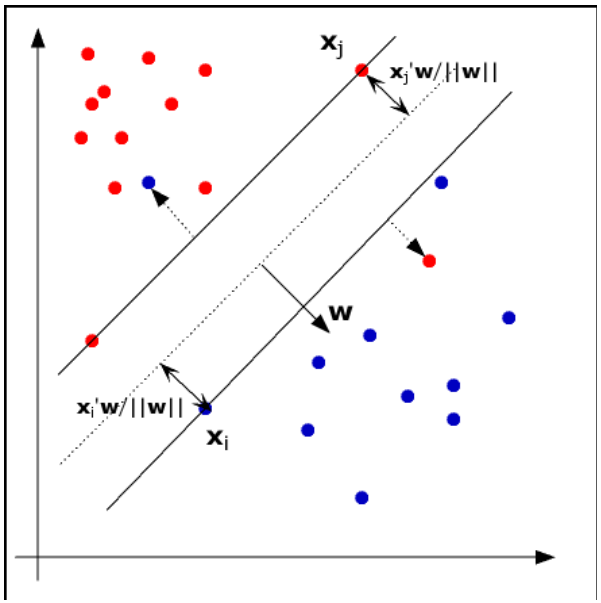
$$\begin{aligned} \left(\mathbf{x}_i'\mathbf{w} + b\right) &\geq 1 \quad \text{if} \quad & y_i &= 1 \\ \left(\mathbf{x}_i'\mathbf{w} + b\right) &\leq -1 \quad \text{if} \quad & y_i &= -1 \end{aligned}$$



- Hence: many $(\mathbf{x}_i, y_i)$ can be separated by a certain margin
- The for which $y_i \left(\mathbf{x}_i'\mathbf{w} + b\right) \leq 1$ are known as the support vectors
- For some, $\left(\mathbf{x}_i'\mathbf{w} + b\right) = y_i$ holds

Support vector machines

- Size of the margin: take a point on the margin, i.e. for which $(\mathbf{x}_i'\mathbf{w} + b) = y_i$, and another point for which $\left(\mathbf{x}_j'\mathbf{w} + b\right) = -1$
- Margin is length of projections of $\mathbf{x}_i$ and $\mathbf{x}_j$ on $\mathbf{w}$: $(\mathbf{x}_i - \mathbf{x}_j)' \mathbf{w} / \|\mathbf{w}\| = 2 / \|\mathbf{w}\|$



Support vector machines

- Capacity functional $\|\mathbf{w}\|^2$ make sure the margin is large...
- At the same time, the pattern function makes sure the classification error on the training set is small...
- The combination of these two features makes sure that the error on another set of data points, a test set, can be expected to be small

Ridge regression: recapitulation

- Optimal $\mathbf{w}$ and b found as:

$$\begin{pmatrix} (\gamma \mathbf{I} + \mathbf{X}'\mathbf{X}) & \mathbf{X}'\mathbf{1} \\ \mathbf{1}'\mathbf{X} & \mathbf{1}'\mathbf{1} \end{pmatrix} \begin{pmatrix} \mathbf{w} \\ b \end{pmatrix} = \begin{pmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{1}'\mathbf{y} \end{pmatrix}$$

- Estimate label for data point $\mathbf{x}$ as $y = \mathbf{x}'\mathbf{w} + b$

Kernel ridge regression

- Note:

$$\left(\mathbf{X}'\mathbf{X} + \gamma\mathbf{I}\right)\mathbf{w} + \mathbf{X}'\mathbf{1}b - \mathbf{X}'\mathbf{y} = 0 \Leftrightarrow \mathbf{w} = \mathbf{X}' \cdot \left(\frac{1}{\gamma}\left(\mathbf{y} - \mathbf{X}\mathbf{w} - \mathbf{1}b\right)\right)$$

- Let's denote $\boldsymbol{\alpha} = \left(\frac{2}{\gamma}\left(\mathbf{y} - \mathbf{X}\mathbf{w} - \mathbf{1}b\right)\right)$, then

$$\mathbf{w} = \mathbf{X}'\boldsymbol{\alpha} = \sum_{i=1}^{n} \alpha_i \mathbf{x}_i$$

- The weight vector is a linear combination of the data points *(representer theorem)*
- Projection of a data point on the weight vector is a weighted sum of kernels (inner products):

$$\mathbf{x}'\mathbf{w} + b = \mathbf{x}'\mathbf{X}'\boldsymbol{\alpha} + b = \sum_{i=1}^{n} \alpha_i k\left(x, x_i\right) + b$$

Kernel ridge regression

Let's plug this in the equations (assuming that $\mathbf{K} = \mathbf{X}\mathbf{X}'$ is full rank):

$$\begin{pmatrix} (\gamma \mathbf{I} + \mathbf{X}'\mathbf{X}) & \mathbf{X}'\mathbf{1} \\ \mathbf{1}'\mathbf{X} & \mathbf{1}'\mathbf{1} \end{pmatrix} \begin{pmatrix} \mathbf{w} \\ b \end{pmatrix} = \begin{pmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{1}'\mathbf{y} \end{pmatrix}$$

$$\begin{pmatrix} \mathbf{X} & \mathbf{0} \\ \mathbf{0} & 1 \end{pmatrix} \begin{pmatrix} (\gamma \mathbf{I} + \mathbf{X}'\mathbf{X}) & \mathbf{X}'\mathbf{1} \\ \mathbf{1}'\mathbf{X} & \mathbf{1}'\mathbf{1} \end{pmatrix} \begin{pmatrix} \mathbf{X}'\boldsymbol{\alpha} \\ b \end{pmatrix} = \begin{pmatrix} \mathbf{X} & \mathbf{0} \\ \mathbf{0} & 1 \end{pmatrix} \begin{pmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{1}'\mathbf{y} \end{pmatrix}$$

$$\begin{pmatrix} \gamma \mathbf{K} + \mathbf{K}^2 & \mathbf{K}\mathbf{1} \\ \mathbf{1}'\mathbf{K} & \mathbf{1}'\mathbf{1} \end{pmatrix} \begin{pmatrix} \boldsymbol{\alpha} \\ b \end{pmatrix} = \begin{pmatrix} \mathbf{K} & \mathbf{0} \\ \mathbf{0} & 1 \end{pmatrix} \begin{pmatrix} \mathbf{y} \\ \mathbf{1}'\mathbf{y} \end{pmatrix}$$

Kernel ridge regression

$$\begin{pmatrix} \gamma\mathbf{K}+\mathbf{K}^2 & \mathbf{K1} \\ \mathbf{1}'\mathbf{K} & \mathbf{1}'\mathbf{1} \end{pmatrix} \begin{pmatrix} \boldsymbol{\alpha} \\ b \end{pmatrix} = \begin{pmatrix} \mathbf{K} & \mathbf{0} \\ \mathbf{0} & 1 \end{pmatrix} \begin{pmatrix} \mathbf{y} \\ \mathbf{1}'\mathbf{y} \end{pmatrix}$$

$$\begin{pmatrix} \gamma\mathbf{I}+\mathbf{K} & \mathbf{1} \\ \mathbf{1}'\mathbf{K} & \mathbf{1}'\mathbf{1} \end{pmatrix} \begin{pmatrix} \boldsymbol{\alpha} \\ b \end{pmatrix} = \begin{pmatrix} \mathbf{y} \\ \mathbf{1}'\mathbf{y} \end{pmatrix}$$

- Again: a set of linear equations...

Kernel ridge regression

- In summary, the dual vector $\boldsymbol{\alpha}$ and the offset b can be found efficiently by solving

$$\begin{pmatrix} \gamma \mathbf{I} + \mathbf{K} & \mathbf{1} \\ \mathbf{1}'\mathbf{K} & \mathbf{1}'\mathbf{1} \end{pmatrix} \begin{pmatrix} \boldsymbol{\alpha} \\ b \end{pmatrix} = \begin{pmatrix} \mathbf{y} \\ \mathbf{1}'\mathbf{y} \end{pmatrix}$$

- Then, for a test object x the label y can be predicted as

$$y = \sum_{i=1}^{n} \alpha_i k(x, x_i) + b$$

Kernel Fisher discriminant analysis

- Just a different use from Kernel ridge regression
- With binary labels y
- $\longrightarrow$ We will not discuss this in greater detail here

Recurring themes and tricks

You should have noticed that all methods relying on inner products, distances,... can be expressed in terms of kernel functions:

1. The 1st step in kernelising invokes an instance of the *representer theorem*: the parameters (weight vector, cluster centre) can be represented as a linear combination of the data:

$$\mathbf{w} = \mathbf{X}\boldsymbol{\alpha}$$

2. The 2nd step plugs in this equation, and *left-multiplies* the equations to obtain inner products $\mathbf{XX}$' where possible...
3. *Kernel trick*: substitute the inner products with kernels

Kernel support vector machines

- Same trick works for support vector machines
- But a different approach is more common here: relying on optimisation theory
- Can be used for ridge regression, PCA, etc as well!

Kernel support vector machines

- Support vector machine:

$$\begin{aligned} \min_{\mathbf{w},b,\xi} \quad & \frac{1}{n}\sum_{i=1}^{n}\xi_i + \gamma \left\|\mathbf{w}\right\|^2 \\ \text{s.t.} \quad & \xi_i \geq 0 \\ & \xi_i \geq 1 - y_i\left(\mathbf{x}_i'\mathbf{w} + b\right) \end{aligned}$$

- Use Lagrange multipliers $\boldsymbol{\alpha} \geq \mathbf{0}$ and $\boldsymbol{\beta} \geq \mathbf{0}$ for both inequalities

Kernel support vector machines

$$\min_{\mathbf{w},b,\boldsymbol{\xi}} \max_{\boldsymbol{\beta}_1,\boldsymbol{\beta}_2} \frac{1}{n}\sum_{i=1}^{n}\xi_i + \gamma\left\|\mathbf{w}\right\|^2 - \boldsymbol{\beta}'\boldsymbol{\xi} - \boldsymbol{\alpha}'\left(\boldsymbol{\xi} - \mathbf{1} + \mathbf{y}\odot\left(\mathbf{X}\mathbf{w} + \mathbf{1}b\right)\right)$$

$$\max_{\boldsymbol{\alpha},\boldsymbol{\beta}} \min_{\mathbf{w},b,\boldsymbol{\xi}} \frac{1}{n}\mathbf{1}'\boldsymbol{\xi} + \gamma\left\|\mathbf{w}\right\|^2 - \left(\boldsymbol{\beta}' + \boldsymbol{\alpha}'\right)\boldsymbol{\xi} + \boldsymbol{\alpha}'\mathbf{1} - \boldsymbol{\alpha}'\mathbf{y}b - \boldsymbol{\alpha}'\left(\mathbf{y}\odot\mathbf{X}\mathbf{w}\right)$$

- Take gradient w.r.t $\mathbf{w}$ and equate to $\mathbf{0}$:

$$2\gamma\mathbf{w} = \mathbf{X}'\,\mathrm{diag}\left(\mathbf{y}\right)\boldsymbol{\alpha}$$

- Same for $\boldsymbol{\xi}$:

$$\frac{1}{n}\mathbf{1} = \left(\boldsymbol{\beta} + \boldsymbol{\alpha}\right)$$

- Same for b:

$$\boldsymbol{\alpha}'\mathbf{y} = \mathbf{0}$$

Kernel support vector machines

- Plugging all this in the objective, gives:

$$\max_{\boldsymbol{\alpha},\boldsymbol{\beta}} \quad -\frac{1}{4\gamma}\boldsymbol{\alpha}'\left(\text{diag}\left(\mathbf{y}\right)\mathbf{X}\mathbf{X}'\,\text{diag}\left(\mathbf{y}\right)\right)\boldsymbol{\alpha}+\boldsymbol{\alpha}'\mathbf{1}$$

$$\max_{\boldsymbol{\alpha},\boldsymbol{\beta}} \quad -\frac{1}{4\gamma}\boldsymbol{\alpha}'\left(\text{diag}\left(\mathbf{y}\right)\mathbf{X}\mathbf{X}'\,\text{diag}\left(\mathbf{y}\right)\right)\boldsymbol{\alpha}+\boldsymbol{\alpha}'\mathbf{1}$$

- Hence, using kernels and with constraints:

$$\begin{aligned}\max_{\boldsymbol{\alpha},\boldsymbol{\beta}} \quad & -\frac{1}{4\gamma}\boldsymbol{\alpha}'\left(\mathbf{K}\odot\ \mathbf{y}\mathbf{y}'\right)\boldsymbol{\alpha}+\boldsymbol{\alpha}'\mathbf{1}\\ \text{s.t.} \quad & \frac{1}{n}\geq\boldsymbol{\alpha}\geq\mathbf{0},\quad \boldsymbol{\alpha}'\mathbf{y}=\mathbf{0}\end{aligned}$$

- This is the Lagrange dual formulation – Lagrange duals are often directly in kernel form...

Averaging pattern functions

- Will follow same pattern as the bound for PCA
- This is due to the fact that both are based on an averaging pattern function
- Let us first do the study in full generality, for averaging pattern functions

$$\pi(X) = \frac{1}{n}\sum_{i=1}^{n} g_\pi(x_i)$$

Averaging pattern functions

- In general:

$$\begin{aligned} \pi(\mathbf{X}) - E_{\underline{\mathbf{X}}}\{\pi(\underline{\mathbf{X}})\} &\leq \max_{\pi\in\Pi}\left(\pi(\mathbf{X}) - E_{\underline{\mathbf{X}}}\{\pi(\underline{\mathbf{X}})\}\right) \\ &\approx E_{\underline{\mathbf{Z}}}\left\{\max_{\pi\in\Pi}\left(\pi(\underline{\mathbf{Z}}) - E_{\underline{\mathbf{X}}}\{\pi(\underline{\mathbf{X}})\}\right)\right\} \\ &\leq E_{\underline{\mathbf{XZ}}}\left\{\max_{\pi\in\Pi}\left(\pi(\underline{\mathbf{Z}}) - \pi(\underline{\mathbf{X}})\right)\right\} \end{aligned}$$

- We should make the approximate inequality into a rigorous inequality...
- Then devise an upper bound for the last quantity

Averaging pattern functions

- The approximate equality for averaging pattern functions:

$$\max_{\pi\in\Pi}\left(\pi\left(\mathbf{X}\right)-E_{\underline{\mathbf{X}}}\left\{\pi\left(\underline{\mathbf{X}}\right)\right\}\right)\approx E_{\underline{\mathbf{Z}}}\left\{\max_{\pi\in\Pi}\left(\pi\left(\underline{\mathbf{Z}}\right)-E_{\underline{\mathbf{X}}}\left\{\pi\left(\underline{\mathbf{X}}\right)\right\}\right)\right\}$$

- Let us assume that $\left|g_{\pi}\left(\mathbf{x}\right)-g_{\pi}\left(\mathbf{x}^{*}\right)\right|\leq M$ (true e.g. if $0\leq g_{\pi}\left(\mathbf{x}\right)\leq M$)
- Then, replacing one data point $\mathbf{x}_i$ by a different value $\mathbf{x}_i^*$ can change the value of this function of $\mathbf{X}$ by at most $\frac{M}{n}$ (this requires some thought... check it!)
- $\rightarrow$ McDiarmid...

- McDiarmid's inequality (again):

Theorem (McDiarmid's inequality)

For f a function of $X = \{x_1, x_2, ..., x_i, ..., x_n\} \in \mathcal{X}$ and $\underline{x_i}$ iid, if $f(X)$ has bounded differences c_i, meaning that $|f(X) - f(X^i)| \leq c_i$, we have that

$$P(f(\underline{X}) - E\{f(\underline{X})\} < \epsilon) \geq 1 - \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^{n} c_i^2}\right).$$

Averaging pattern functions

- McDiarmid's inequality with $c_i = \frac{M}{n}$: with probability at least $1 - \exp\left(-\frac{2n\epsilon^2}{M^2}\right)$

$$\max_{\pi\in\Pi}\left(\pi\left(\mathbf{X}\right) - E_{\underline{\mathbf{X}}}\left\{\pi\left(\underline{\mathbf{X}}\right)\right\}\right) - E_{\underline{\mathbf{Z}}}\left\{\max_{\pi\in\Pi}\left(\pi\left(\underline{\mathbf{Z}}\right) - E_{\underline{\mathbf{X}}}\left\{\pi\left(\underline{\mathbf{X}}\right)\right\}\right)\right\} < \epsilon$$

- In other words, with a probability of at least $\delta/2$, we have that:

$$\begin{aligned} & \max_{\pi\in\Pi}\left(\pi\left(\mathbf{X}\right) - E_{\underline{X}}\left\{\pi\left(\underline{\mathbf{X}}\right)\right\}\right) \\ \leq \quad & E_{\underline{\mathbf{Z}}}\left\{\max_{\pi\in\Pi}\left(\pi\left(\underline{\mathbf{Z}}\right) - E_{\underline{\mathbf{X}}}\left\{\pi\left(\underline{\mathbf{X}}\right)\right\}\right)\right\} + M\sqrt{\frac{\ln\left(2/\delta\right)}{2n}}. \end{aligned}$$

Averaging pattern functions

- $\Rightarrow$ quantity to be bounded: $E_{\underline{\mathbf{XZ}}}\left\{\max_{\pi\in\Pi}\left(\pi\left(\underline{\mathbf{Z}}\right)-\pi\left(\underline{\mathbf{X}}\right)\right)\right\}$
- Bounded by the Rademacher complexity (σ_i i.i.d., 1 or -1 both with probability $\frac{1}{2}$)

$$
\begin{aligned}
& E_{\underline{\mathbf{XZ}}}\left\{\max_{\pi}\left(\pi\left(\underline{\mathbf{Z}}\right)-\pi\left(\underline{\mathbf{X}}\right)\right)\right\} \\
= \quad & E_{\underline{\mathbf{XZ}}}\left\{\max_{\pi\in\Pi}\left(\frac{1}{n}\sum_{i=1}^{n}\left(g_\pi\left(\underline{\mathbf{z}}_i\right)-g_\pi\left(\underline{\mathbf{x}}_i\right)\right)\right)\right\} \\
= \quad & E_{\underline{\mathbf{XZ}}\sigma}\left\{\max_{\pi\in\Pi}\left|\frac{1}{n}\sum_{i=1}^{n}\underline{\sigma_i}\left(g_\pi\left(\underline{\mathbf{z}}_i\right)-g_\pi\left(\underline{\mathbf{x}}_i\right)\right)\right|\right\} \\
\leq \quad & E_{\underline{\mathbf{X}}\sigma}\left\{\max_{\pi\in\Pi}\left|\frac{2}{n}\sum_{i=1}^{n}\underline{\sigma_i}g_\pi\left(\underline{\mathbf{x}}_i\right)\right|\right\} \triangleq \mathcal{R}\left(\Pi\right),
\end{aligned}
$$

Rademacher complexity

Definition (Rademacher and empirical Rademacher complexity)

The *Rademacher complexity* $\mathcal{R}(\Pi)$ of a space Π of additive pattern functions π with $\pi(X) = \frac{1}{n}\sum_{i=1}^{n} g(\mathbf{x}_i)$ is given by

$$\mathcal{R}(\Pi) = E_{\underline{\mathbf{X}}\underline{\sigma}}\left\{\max_{\pi\in\Pi}\left|\frac{2}{n}\sum_{i=1}^{n}\underline{\sigma_i} g_\pi\left(\underline{\mathbf{x}_i}\right)\right|\right\}.$$

The *empirical Rademacher complexity* $\widehat{\mathcal{R}}_{\mathbf{X}}(\Pi)$ of the same pattern space and for given data $X = \{x_1, x_2, \ldots, x_n\}$ is given by

$$\widehat{\mathcal{R}}_{\mathbf{X}}(\Pi) = E_{\underline{\sigma}}\left\{\max_{\pi\in\Pi}\left|\frac{2}{n}\sum_{i=1}^{n}\underline{\sigma_i} g_\pi\left(\mathbf{x}_i\right)\right|\right\}.$$

Rademacher complexity

- McDiarmid's inequality $\Rightarrow \widehat{\mathcal{R}}_{\mathbf{X}}(\Pi) \approx \mathcal{R}(\Pi)$ with high probability
- Indeed, it is easy to very that McDiarmid's theorem applies with $c_i = \frac{2M}{n}$, showing that with a probability of at least $\delta/2$

$$\mathcal{R}(\Pi) \leq \widehat{\mathcal{R}}_{\mathbf{X}}(\Pi) + 2M\sqrt{\frac{\ln(2/\delta)}{2n}}$$

Rademacher bounds

- The resulting empirical Rademacher type bound is given by

$$\begin{aligned} \pi(\mathbf{X}) - E_{\underline{\mathbf{X}}}\{\pi(\underline{\mathbf{X}})\} &= \pi(\mathbf{X}) - E_{\underline{\mathbf{x}}}\{g_\pi(\underline{\mathbf{x}})\} \\ &\leq \widehat{\mathcal{R}}_{\mathbf{X}}(\Pi) + 3M\sqrt{\frac{\ln(2/\delta)}{2n}} \end{aligned}$$

  which holds with probability $1 - 2 \cdot \delta/2 = 1 - \delta$ over random draws of $\mathbf{X}$
- Hereby, M is an upper bound on $|g_\pi(\mathbf{x}) - g_\pi(\mathbf{x}^*)| \quad (\forall \mathbf{x}, \mathbf{x}^*)$
- Power of this type of bounds:
  - Quite tight, data-dependent
  - $\widehat{\mathcal{R}}_{\mathbf{X}}(\Pi)$ is usually easy to bound

Ridge regression stability bound (without offset)

- We prove stability for the stratification formulation:

$$\min_{\mathbf{w}} \frac{1}{n} ||\mathbf{X}\mathbf{w} - \mathbf{y}||^2 \quad \text{s.t.} \quad ||\mathbf{w}||^2 \leq c$$

- Assume: $\|\mathbf{x}\|^2 \leq R_{\mathbf{x}}^2$ and $y \leq R_y$
- $0 \leq g_{\pi_w}(\mathbf{x}, y) = (\mathbf{x}'\mathbf{w} - y)^2 \leq cR_{\mathbf{x}}^2 + 2\sqrt{c}R_x R_y + R_y^2$, so: $M = cR_{\mathbf{x}}^2 + 2\sqrt{c}R_x R_y + R_y^2$
- Empirical Rademacher complexity:

$$\begin{aligned} & \widehat{\mathcal{R}}_{\mathbf{X}}(\Pi) \\ \leq \quad & \frac{2}{n} c \sqrt{\sum_{i=1}^{n} (\mathbf{x}_i'\mathbf{x}_i)^2} + \frac{2}{n} \sqrt{\sum_{i=1}^{n} y_i^4} + \frac{4}{n} \sqrt{c} \sqrt{\sum_{i=1}^{n} y_i^2 (\mathbf{x}_i'\mathbf{x}_i)}. \end{aligned}$$

Ridge regression stability bound (without offset)

$$
\begin{aligned}
\widehat{\mathcal{R}}_{\mathbf{X}}(\Pi) &= E_{\underline{\sigma}}\left\{\max_{\mathbf{w}}\left|\frac{2}{n}\sum_{i=1}^{n}\underline{\sigma_i}\left(\mathbf{x}_i'\mathbf{w}-y_i\right)^2\right|\right\} \\
&= E_{\underline{\sigma}}\left\{\max_{\mathbf{w}}\left|\frac{2}{n}\sum_{i=1}^{n}\underline{\sigma_i}\left(\left(\mathbf{x}_i'\mathbf{w}\right)^2+y_i^2-2y_i\mathbf{x}_i'\mathbf{w}\right)\right|\right\} \\
&\leq \frac{2}{n}E_{\underline{\sigma}}\left\{\max_{\mathbf{w}}\left(\left|\sum_{i=1}^{n}\underline{\sigma_i}\left(\mathbf{x}_i'\mathbf{w}\right)^2\right|+\left|\sum_{i=1}^{n}\underline{\sigma_i}y_i^2\right|+2\left|\sum_{i=1}^{n}\underline{\sigma_i}y_i\mathbf{x}_i'\mathbf{w}\right|\right)\right. \\
&\leq \frac{2}{n}E_{\underline{\sigma}}\left\{\begin{array}{c}\max_{\mathbf{w}}\left|\sum_{i=1}^{n}\left\langle\underline{\sigma_i}\mathbf{x}_i\mathbf{x}_i',\mathbf{w}\mathbf{w}'\right\rangle\right|+\max_{\mathbf{w}}\left|\sum_{i=1}^{n}\underline{\sigma_i}y_i^2\right| \\ +2\max_{\mathbf{w}}\left|\sum_{i=1}^{n}\left\langle\underline{\sigma_i}y_i\mathbf{x}_i,\mathbf{w}\right\rangle\right|\end{array}\right\}
\end{aligned}
$$

Ridge regression stability bound (without offset)

$$\widehat{\mathcal{R}}_{\mathbf{x}}(\Pi) \leq \frac{2}{n} E_{\underline{\sigma}} \left\{ \begin{array}{c} \sqrt{c^2} \sqrt{\sum_{i,j=1}^{n} \left\langle \underline{\sigma_i} \mathbf{x}_i \mathbf{x}_i', \underline{\sigma_j} \mathbf{x}_j \mathbf{x}_j' \right\rangle} \\ + \sqrt{\sum_{i,j=1}^{n} \underline{\sigma_i} \underline{\sigma_j} y_i^2 y_j^2} \\ + 2\sqrt{c} \sqrt{\sum_{i,j=1}^{n} \left\langle \underline{\sigma_i} y_i \mathbf{x}_i, \underline{\sigma_j} y_j \mathbf{x}_j \right\rangle} \end{array} \right\}$$

$$\leq \frac{2}{n} c \sqrt{E_{\underline{\sigma}} \left\{ \sum_{i,j=1}^{n} \underline{\sigma_i} \underline{\sigma_j} \left\langle \mathbf{x}_i \mathbf{x}_i', \mathbf{x}_j \mathbf{x}_j' \right\rangle \right\}} + \frac{2}{n} \sqrt{E_{\underline{\sigma}} \left\{ \sum_{i,j=1}^{n} \underline{\sigma_i} \underline{\sigma_j} y_i^2 y_j^2 \right\}}$$

$$+ \frac{4}{n} \sqrt{c} \sqrt{E_{\underline{\sigma}} \left\{ \sum_{i,j=1}^{n} \underline{\sigma_i} \underline{\sigma_j} \left\langle y_i \mathbf{x}_i, y_j \mathbf{x}_j \right\rangle \right\}}$$

Ridge regression stability bound (without offset)

$$
\begin{aligned}
\widehat{\mathcal{R}}_{\mathbf{X}}(\Pi) \quad \leq \quad & \frac{2}{n} c \sqrt{E_{\underline{\sigma}} \left\{ \sum_{i,j=1}^{n} \underline{\sigma_i} \underline{\sigma_j} \left\langle \mathbf{x}_i \mathbf{x}_i', \mathbf{x}_j \mathbf{x}_j' \right\rangle \right\}} \\
& + \frac{2}{n} \sqrt{E_{\underline{\sigma}} \left\{ \sum_{i,j=1}^{n} \underline{\sigma_i} \underline{\sigma_j} y_i^2 y_j^2 \right\}} \\
& + \frac{4}{n} \sqrt{c} \sqrt{E_{\underline{\sigma}} \left\{ \sum_{i,j=1}^{n} \underline{\sigma_i} \underline{\sigma_j} \left\langle y_i \mathbf{x}_i, y_j \mathbf{x}_j \right\rangle \right\}} \\
\leq \quad & \frac{2}{n} c \sqrt{\sum_{i=1}^{n} \left(\mathbf{x}_i' \mathbf{x}_i \right)^2} + \frac{2}{n} \sqrt{\sum_{i=1}^{n} y_i^4} + \frac{4}{n} \sqrt{c} \sqrt{\sum_{i=1}^{n} y_i^2 \left(\mathbf{x}_i' \mathbf{x}_i \right)}
\end{aligned}
$$

Wrap-up

- Supervised learning methods:
  - Ridge regression revisited (now with offset)
  - Fisher's discriminant analysis
  - Support vector machine
- Kernel versions
- Statistical study
  - In general for averaging pattern functions using Rademacher complexities
  - In particular, applied to ridge regression