Kernels

We might want to consider something more complicated than a linear model:

**Example 1:** \([x^{(1)}, x^{(2)}] \rightarrow \Phi \left([x^{(1)}, x^{(2)}]\right) = [x^{(1)2}, x^{(2)2}, x^{(1)}x^{(2)}]\)

Information unchanged, but now we have a **linear** classifier on the transformed points.

With the kernel trick, we just need kernel \(k(a, b) = \Phi(a)^T \Phi(b)\)
Example 4:

\[ k(x, z) = (x^T z + c)^2 = \left( \sum_{j=1}^{n} x^{(j)} z^{(j)} + c \right) \left( \sum_{\ell=1}^{n} x^{(\ell)} z^{(\ell)} + c \right) \]

\[ = \sum_{j=1}^{n} \sum_{\ell=1}^{n} x^{(j)} x^{(\ell)} z^{(j)} z^{(\ell)} + 2c \sum_{j=1}^{n} x^{(j)} z^{(j)} + c^2 \]

\[ = \sum_{j,\ell=1}^{n} (x^{(j)} x^{(\ell)})(z^{(j)} z^{(\ell)}) + \sum_{j=1}^{n} (\sqrt{2c} x^{(j)})(\sqrt{2c} z^{(j)}) + c^2, \]

and in \( n = 3 \) dimensions, one possible feature map is:

\[ \Phi(x) = [x^{(1)^2}, x^{(1)} x^{(2)}, \ldots, x^{(3)^2}, \sqrt{2cx^{(1)}}, \sqrt{2cx^{(2)}}, \sqrt{2cx^{(3)}}, c] \]

and \( c \) controls the relative weight of the linear and quadratic terms in the inner product.

Even more generally, if you wanted to, you could choose the kernel to be any higher power of the regular inner product.
Dual representation, Sec 6.2

Primal problem: \( \min_w E(w) \)

\[
E = \frac{1}{2} \sum_n \left\{ w^T x_n - t_n \right\}^2 + \frac{\lambda}{2} \|w\|^2 = \|Xw - t\|_2^2 + \frac{\lambda}{2} \|w\|^2
\]

Solution \( w = X^+ t = (X^T X + \lambda I_M)^{-1} X^T t \)

\[
= X^T (XX^T + \lambda I_N)^{-1} t = X^T (K + \lambda I_N)^{-1} t = X^T a
\]

The kernel is \( K = XX^T \)

Dual representation is : \( \min_a E(a) \)

\[
E = \frac{1}{2} \sum_n \left\{ w^T x_n - t_n \right\}^2 + \frac{\lambda}{2} \|w\|^2 = \|Ka - t\|_2^2 + \frac{\lambda}{2} a^T Ka
\]

Prediction

\[
y = w^T x = a^T Xx = \sum_n a_n x_n^T x = \sum_n a_n k(x_n, x)
\]
Dual representation, Sec 6.2

Prediction

\[ y = w^T x = a^T X x = \sum_{n} a_n x_n^T x = \sum_{n} a_n k(x_n, x) \]

- Often \textbf{a is sparse} (... Support vector machines)
- We don’t need to know \( x \) or \( \varphi(x) \). \textit{Just the Kernel}

\[
E(a) = \|Ka - t\|^2 + \frac{\lambda}{2} a^T Ka
\]
Gaussian Kernels

- Gaussian Kernel

\[ k(x, x') = \exp \left( -\frac{1}{2} (x - x')^T \Sigma^{-1} (x - x') \right) \]

Diagonal \( \Sigma \): (this gives ARD)

\[ k(x, x') = \exp \left( -\frac{1}{2} \sum_{i=1}^{N} \frac{(x_i - x'_i)^2}{\sigma_i^2} \right) \]

Isotropic \( \sigma_i^2 \) gives an RBF

\[ k(x, x') = \exp \left( -\frac{\|x - x'\|_2^2}{2\sigma^2} \right) \]
Can be inner product in infinite dimensional space

Assume \( x \in \mathbb{R}^1 \) and \( \gamma > 0 \).

\[
e^{-\gamma \|x_i - x_j\|^2} = e^{-\gamma (x_i - x_j)^2} = e^{-\gamma x_i^2 + 2\gamma x_i x_j - \gamma x_j^2}
\]

\[
e^{-\gamma x_i^2 - \gamma x_j^2} \left( 1 + \frac{2\gamma x_i x_j}{1!} + \frac{(2\gamma x_i x_j)^2}{2!} + \frac{(2\gamma x_i x_j)^3}{3!} + \cdots \right)
\]

\[
e^{-\gamma x_i^2 - \gamma x_j^2} \left( 1 \cdot 1 + \sqrt{\frac{2\gamma}{1!}} x_i \cdot \sqrt{\frac{2\gamma}{1!}} x_j + \sqrt{\frac{(2\gamma)^2}{2!}} x_i^2 \cdot \sqrt{\frac{(2\gamma)^2}{2!}} x_j^2 
\]

\[
+ \sqrt{\frac{(2\gamma)^3}{3!}} x_i^3 \cdot \sqrt{\frac{(2\gamma)^3}{3!}} x_j^3 + \cdots \right) = \phi(x_i)^T \phi(x_j),
\]

where

\[
\phi(x) = e^{-\gamma x^2} \left[ 1, \sqrt{\frac{2\gamma}{1!}} x, \sqrt{\frac{(2\gamma)^2}{2!}} x^2, \sqrt{\frac{(2\gamma)^3}{3!}} x^3, \cdots \right]^T.
\]
Sparse Bayesian Learning (SBL)

Model: \( y = Ax + n \)
Prior: \( x \sim \mathcal{N}(x; 0, \Gamma) \)
\( \Gamma = \text{diag}(\gamma_1, \ldots, \gamma_M) \)
Likelihood: \( p(y|x) = \mathcal{N}(y; Ax, \sigma^2 I_N) \)

Evidence: \( p(y) = \int_x p(y|x)p(x)dx = \mathcal{N}(y; 0, \Sigma_y) \)
\( \Sigma_y = \sigma^2 I_N + A\Gamma A^H \)

SBL solution: \( \hat{\Gamma} = \arg \max_\Gamma p(y) \)
\[ = \arg \min_\Gamma \left\{ \log |\Sigma_y| + y^H \Sigma_y^{-1} y \right\} \]

Lecture 10
Support Vector Machines

Non Bayesian!

Features:
• Kernel
• Sparse representations
• Large margins
Regularize for plausibility

- Which one is best?
- We maximize the margin
Regularize for plausibility
Support Vector Machines

• The line that maximizes the minimum margin is a good bet.
  – The model class of “hyper-planes with a margin $m$” has a low VC dimension if $m$ is big.

• This maximum-margin separator is determined by a subset of the datapoints.
  – Datapoints in this subset are called “support vectors”.
  – It is useful computationally if only few datapoints are support vectors, because the support vectors decide which side of the separator a test case is on.

The support vectors are indicated by the circles around them.
Lagrange multiplier (Bishop App E)

\[
\max(f(x)) \text{ subject to } g(x) = 0
\]

Taylor expansion
\[
g(x + \varepsilon) = g(x) + \varepsilon^T \nabla g(x)
\]

\[
L(x, \lambda) = f(x) + \lambda g(x)
\]
Lagrange multiplier (Bishop App E)

\[
\max(f(x)) \text{ subject to } g(x) > 0
\]
\[
L(x, \lambda) = f(x) + \lambda g(x)
\]

Either \( \nabla f(x) = 0 \)
Then \( g(x) \) is inactive, \( \lambda = 0 \)

Or \( g(x) = 0 \) but \( \lambda > 0 \)

Thus optimizing \( L(x, \lambda) \) with the Karesh-Kuhn-Trucker (KKT) equations

\[
g(x) \geq 0
\]
\[
\lambda \geq 0
\]
\[
\lambda g(x) = 0
\]
Testing a linear SVM

- The separator is defined as the set of points for which:

\[ \mathbf{w} \cdot \mathbf{x} + b = 0 \]

so if \( \mathbf{w} \cdot \mathbf{x}^c + b > 0 \) say its a positive case

and if \( \mathbf{w} \cdot \mathbf{x}^c + b < 0 \) say its a negative case
Discriminant functions

The planar decision surface in data-space for the simple linear discriminant function:

$$ w^T x + w_0 \geq 0 $$

$$ y = w^T x + w_0 $$

$$ y_{\perp} = w^T x_{\perp} + w_0 = 0 $$

$$ \|w\|^2 = w^T w $$

X on plane => y=0 =>

Distance from plane

$$ r = \frac{y}{\|w\|^2} $$

$$ X = X_{\perp} + r \frac{w_0}{\|w\|^2} $$

$$ w^T x = w^T x_{\perp} + r \frac{w^T w}{\|w\|^2} $$

$$ X = X_{\perp} + r \frac{w_0}{\|w\|^2} $$

$$ X = X_{\perp} + r \frac{y}{\|w\|^2} $$
Large margin

\[ y = w^T \phi(x) + b \]

\[ x_n = x_{\perp} + r_n \frac{w}{\|w\|} \]

\( x \) on plane \( \Rightarrow y=0 \) \( \Rightarrow b = -w^T \phi(x) \)

\[ r_n = \frac{w^T \phi(x_n) + b}{\|w\|} = \frac{y_n}{\|w\|} \]

\[ t_n y_n \geq 1 \]

\[ \max_n \frac{1}{\|w\|} \min_n t_n y_n \]
Maximum margin (Bishop 7.1)

\[
\arg\min_{w,b} \frac{1}{2} \|w\|^2
\]

Subject to

\[
t_n \left( w^T \phi(x_n) + b \right) \geq 1, \quad n = 1, \ldots, N.
\] (7.5)

Lagrange function

\[
L(w, b, a) = \frac{1}{2} \|w\|^2 - \sum_{n=1}^{N} a_n \left\{ t_n (w^T \phi(x_n) + b) - 1 \right\}
\] (7.7)

Differentiation

\[
w = \sum_{n=1}^{N} a_n t_n \phi(x_n)
\] (7.8)

\[
0 = \sum_{n=1}^{N} a_n t_n.
\] (7.9)

Dual representation

\[
\tilde{L}(a) = \sum_{n=1}^{N} a_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} a_n a_m t_n t_m k(x_n, x_m)
\] (7.10)

with respect to \( a \) subject to the constraints

\[
a_n \geq 0, \quad n = 1, \ldots, N,
\] (7.11)

\[
\sum_{n=1}^{N} a_n t_n = 0.
\] (7.12)

This can be solved with quadratic programming
Maximum margin (Bishop 7.1)

• KKT conditions

\[ a_n \geq 0 \quad (7.14) \]
\[ t_n y(x_n) - 1 \geq 0 \quad (7.15) \]
\[ a_n \{ t_n y(x_n) - 1 \} = 0. \quad (7.16) \]

either \( a_n = 0 \) or \( t_n y(x_n) = 1 \).

• Solving for \( a_n \)

\[ w = \sum_{n=1}^{N} a_n t_n \phi(x_n) \quad (7.8) \]

• Prediction

\[ y(x) = \sum_{n=1}^{N} a_n t_n k(x, x_n) + b. \quad (7.13) \]
If there is no separating plane…

- Use a bigger set of features.
  - Makes the computation slow? “Kernel” trick makes the computation fast with many features.

- Extend definition of maximum margin to allow non-separating planes.
  - Use “slack” variables \( \xi = |t_n - y(x_n)| \)

\[
t_n y(x_n) \geq 1 - \xi_n, \quad n = 1, \ldots, N
\]  
(7.20)

Objective function

\[
C \sum_{n=1}^{N} \xi_n + \frac{1}{2}||w||^2
\]  
(7.21)
SVM classification summarized--- Only kernels

- Minimize with respect to $\mathbf{w}, w_0$

\[
C \sum_n^{N} \zeta_n + \frac{1}{2} \|\mathbf{w}\|^2
\]  
(Bishop 7.21)

- Solution found in dual domain with Lagrange multipliers
  - $a_n, n = 1 \cdots N$ and

- This gives the support vectors $S$

\[
\hat{\mathbf{w}} = \sum_{n \in S} a_n t_n \varphi(x_n)
\]  
(Bishop 7.8)

- Used for predictions

\[
\hat{y} = w_0 + \mathbf{w}^T \varphi(x) = w_0 + \sum_{n \in S} a_n t_n \varphi(x_n)^T \varphi(x)
\]

\[
= w_0 + \sum_{n \in S} a_n t_n k(x_n, x)
\]  
(Bishop 7.13)
14.5.1 SVMs for regression

The problem with kernelized ridge regression is that the solution vector in the structured output case; see Section 19.7.

However, we include a discussion of SVMs, despite their non-probabilistic nature, for two main reasons. First, they are very popular and widely used, so all students of machine learning should know about them. Second, they have some computational advantages over probabilistic methods. Despite these advantages, SVMs are not as versatile as probabilistic methods. For example, they are not as well suited for multi-class classification as we explain below.

It is possible to obtain sparse, probabilistic, multi-class kernel-based classifiers, which work as well or better than SVMs, using techniques such as the L1VM or RVM, discussed in Section 14.3.2. For more details, see (Vapnik et al. 1997).

Figure 14.10. (a) Illustration of $\ell_2$, Huber and $\epsilon$-insensitive loss functions, where $\epsilon = 1.5$. Figure generated by huberLossDemo. (b) Illustration of the $\epsilon$-tube used in SVM regression. Points above the tube have $\xi_i > 0$ and $\xi_i^* = 0$. Points below the tube have $\xi_i = 0$ and $\xi_i^* > 0$. Points inside the tube have $\hat{\xi}_i = \xi_i = 0$. Based on Figure 7.7 of (Bishop 2006a).
SVMs are Perceptrons!

• SVM’s use each training case, \( x \), to define a feature \( K(x, .) \) where \( K \) is user chosen.
  – So the user designs the features.
• SVM do “feature selection” by picking support vectors, and learn feature weighting from a big optimization problem.
• => SVM is a clever way to train a standard perceptron.
  – What a perceptron cannot do, SVM cannot do.

• SVM DOES:
  – Margin maximization
  – Kernel trick
  – Sparse
SVM Code for classification (libsvm)

Part of ocean acoustic data set http://noiselab.ucsd.edu/ECE285/SIO209Final.zip case 'Classify'

% train
model = svmtrain(Y, X,['-c 7.46 -g ' gamma ' -q ' kernel]);

% predict
[predict_label,~, ~] = svmpredict(rand([length(Y),1]), X, model,'-q');

>> model = struct with fields:
   Parameters: [5x1 double]
   nr_class: 2
   totalSV: 36
   rho: 8.3220
   Label: [2x1 double]
   sv_indices: [36x1 double]
   ProbA: []  ProbB: []
   nSV: [2x1 double]
   sv_coef: [36x1 double]
   SVs: [36x2 double]
Finding the Decision Function

- **w**: maybe infinite variables
- The dual problem

\[
\begin{align*}
\min_{\alpha} & \quad \frac{1}{2} \alpha^T Q \alpha - e^T \alpha \\
\text{subject to} & \quad 0 \leq \alpha_i \leq C, \ i = 1, \ldots, l \\
& \quad y^T \alpha = 0,
\end{align*}
\]

where \( Q_{ij} = y_i y_j \phi(x_i)^T \phi(x_j) \) and \( e = [1, \ldots, 1]^T \)

- At optimum

\[
w = \sum_{i=1}^l \alpha_i y_i \phi(x_i)
\]

- A finite problem: \#variables = \#training data

Using these results to eliminate \( w, b, \) and \( \{\xi_n\} \) from the Lagrangian, we obtain the dual Lagrangian in the form

\[
\tilde{L}(\alpha) = \sum_{n=1}^N a_n - \frac{1}{2} \sum_{n=1}^N \sum_{m=1}^N a_n a_m t_n t_m k(x_n, x_m)
\]

(7.32)

Corresponds to (Bishop 7.32) With \( y=t \)
1. Fitting the spiral with default settings fail due to the small training set. The NN will fit to the training data which is not representative of the true pattern and the network will **generalize** poorly. Increasing the ratio of training to test data to 90% the NN finds the correct shape (1\textsuperscript{st} image).
Tensorflow Playground

You can fix the generalization problem by adding **noise** to the data. This allows the small training set to generalize better as it reduces **overfitting** of the training data (2nd image).
Adding an additional hidden layer the NN fails to classify the shape properly. **Overfitting** once again becomes a problem even after you've added noise. This can be fixed by adding appropriate **L2 regularization** (third image).
• NOT USED