Logistic Regression
Support Vector Machines

- Pick up assignment 1 at end of class
- Assignment 2 due now
- Assignment 3 out (due Wednesday)
- Tutorials are now 3-4:30pm Thursdays (FOM) 3:30-4:30pm Fridays (IC)
- Comments on lecture style (context), course eval needed
  - material
  - chol group - L, reg
Why do we need a valid kernel? (k > 0)

1. Interpretation
   - Valid kernels correspond to some feature space
     \[ K(x, x') = \phi(x) \phi(x') \]
   - For Gaussian, valid kernels ensure likelihood is normalized

2. Convexity (Monday)
   - Can't solve efficiently

But, non-valid kernels may still perform well.
Big picture of supervised learning

- Generative model
  - MLE $p(y_i, x_i | \theta)$
  - MAP $p(y_i, x_i, \theta) / p(\theta)$
  - logistic regression
- Discriminative model
  - Discriminant function $f(x_i, \theta) \rightarrow y_i$
  - Support vector machines

Today: linear discriminative models, discriminant functions
Test line: You get an e-mail from your bank.

"bad errors"

Training: Two "number of errors" of grammar:

\[ y = \begin{cases} 1 & \text{if } x = \text{error} \\ 0 & \text{otherwise} \end{cases} \]

on \( w \)
Logistic Regression

Support Vector Machines

Logistic Regression

(parametric) linear, better handling of "bad errors"

\[ p(y_i = 1 | x_i, \omega) = \frac{1}{1 + \exp(-\omega^T x_i)} \]

\[ p(y_i = -1 | x_i, \omega) = \frac{1}{1 + \exp(\omega^T x_i)} \]

Sigmoid \( \sigma \)
\[ \sum_{i=1}^{N} -\log p(y_i | \bar{x}_i; \bar{w}) = \sum_{i=1}^{N} \log (1 + \exp(-y_i \bar{w}^T \bar{x}_i)) \]

\[ \frac{1}{2} (\bar{w}^T \bar{x}_i - y_i)^2 \]

\[ \log (1 + \exp(\beta)) \]

\[ \beta \]

\[ \beta = 1 \]

\[ \beta = -1 \]
MAP logistic

\[ \log \text{-likelihood} \]

\[ \sum \log (1 + \exp(-x_i \hat{w}^T \bar{w})) + \frac{\lambda}{2} \| \hat{w} \|_2^2 \]

Can kernelize
Multinomial logistic regression

\[ x_i \in \mathbb{R}^d \]
\[ y_i \in \{1, 2, 3, \ldots, K\} \]

"Tony, Jerry, "Etch", Scratchy"

We'll have \( \vec{w}_c \) for each class \( c \).

\[
p(y_i = c | x_i) = \frac{\exp(\vec{w}_c^T x_i)}{\sum_{c=1}^{K} \exp(\vec{w}_c^T x_i)}
\]

"softmax"

"Over-parameterized"

\[
\sum_{i=1}^{n} p(y_i = c | x_i; \vec{w}) = 1
\]

MLE: set \( \vec{w}_c = 0 \) for 1 class \( c \).
\[
\sum_{i=1}^{N} -\log p(y_i | \bar{w}, x_i) = -\log \left( \sum_{c=1}^{K} \frac{\exp(\bar{w}_{y_i}^T x_i)}{\sum_{c=1}^{K} \exp(\bar{w}_c^T x_i)} \right)
\]

\[
= \sum_{i=1}^{N} -\bar{w}_{y_i}^T x_i + \log \left( \sum_{c=1}^{K} \frac{\exp(\bar{w}_c^T x_i)}{\sum_{c=1}^{K} \exp(\bar{w}_c^T x_i)} \right)
\]

\[
\sum_{i=1}^{N} -I(c=y_i) \bar{w}_c^T x_i + \frac{\exp(\bar{w}_c^T x_i)}{\sum_{c=1}^{K} \exp(\bar{w}_c^T x_i)} x_i
\]

\[
p(y_i = c | x_i, \bar{w})
\]
**Log-sum-exp trick**

\[
\log \left( \sum_{i=1}^{N} \exp(a_i) \right) = \log \left( \sum_{i=1}^{N} \exp(a_i - A + A) \right), \quad A
\]

\[
= \log \left( \exp(A) \sum_{i=1}^{N} \exp(a_i - A) \right) \\
= A + \log \left( \sum_{i=1}^{N} \exp(a_i - A) \right)
\]
Logistic Regression

Support Vector Machines

Support Vector Machines

\[ \min_{\mathbf{w}} \frac{1}{2} \mathbf{w}^T \mathbf{w} \]

\[ \text{s.t. } \mathbf{w}^T \mathbf{x}_i \geq 1, \quad y_i = 1 \]

\[ \mathbf{w}^T \mathbf{x}_i \leq -1, \quad y_i = -1 \]

\[ \mathbf{w}^T \mathbf{x} = 0 \]