MLE/MAP \( \log \text{variance} \)

Support Vector Machines

Convex functions

"Standardize Features"

- When do MAP, we often standardize columns

\[
X = \begin{bmatrix}
    X_1 \\
    \vdots \\
    X_n
\end{bmatrix} \quad \text{In regression we sometimes do this to } Y
\]

so we don't need y-intercept.

- Subtract mean \( \mu \), mean 0
- Divide by variance \( \sigma \), variance 1
- Roughly put weight \( w_2 \) on same side.

"Bias" vs "bias"

- Sometimes you'll see logistic regression written using

\[
p(y=1 | x, \theta) = \frac{1}{1 + \exp(-\theta^T x)}
\]

instead of

\[
p(y=1 | x, \theta) = \frac{1}{1 + \exp(-\theta^T x)}
\]

- "Bias" variable \( b \) takes into account that
- one class may be more likely before we see the features
- (e.g., \( p(y=1) \) in naive Bayes)
- Equivalent to adding a column of ones to \( X \)
- Statisticians say: "Do not regularize the bias"

Linear Classifiers

- Generative
- Discriminative

<table>
<thead>
<tr>
<th>Naive Bayes</th>
<th>Logistic Regression</th>
<th>Support Vector Machines</th>
</tr>
</thead>
</table>

Support Vector Machines

- \# training requests
- \# grammar errors

Linear classifier

- Linear classifier (linear classifier) hold

"Linearly separable"
Maximum Margin

To find the optimal line:

"margin" is proportional to $\frac{1}{||w||}$

\[ \text{maximize } \frac{1}{||w||} \iff \text{minimizing } ||w||^2 \]

SVM:

\[ \min_{w} \frac{1}{2} ||w||^2 \quad \text{s.t.} \quad y_i (w^T x_i - 1) \geq 0 \]

Find best $w$ such that:

- $0$ points on one side
- $1$ points on other side

when $w$ is equal (when $w$ is not support vectors) $w^T x_i \geq 1 \quad y_i = 1$

"Soft margin" SVM

\[ \min_{\alpha, \nu} \sum_{i=1}^{N} \nu_i + \frac{1}{2} ||w||^2 \quad \text{s.t.} \quad y_i (w^T x_i - 1) + \nu_i \geq 0 \]

let $\nu_i$ violate the constraint a bit

for $C$ large enough, linearly separable is a special case

\[ \nu_i \geq \max \left\{ 0, 1 - y_i w^T x_i \right\} \]

\[ \text{argmin}_{w} \sum_{i=1}^{N} \max \left\{ 0, 1 - y_i (w^T x_i) \right\} + \frac{1}{2} ||w||^2 \]

"hinge loss"

\[ \text{linear function} \quad \text{max is zero} \]

Multi-Class SVM

\[ \text{e.g. } y_i \in \{1, 2, 3 \ldots K \} \]

or $y_i \in \{ \text{red, blue, \ldots} \}$

- 1 vs all:
  - train $K$ binary classifiers $\text{O}(K)$
  - vs all
- all pairs:
  - 1 vs 2
  - 2 vs 3
- EOC: error correcting output code
- Multiclass:
  \[
  \min_{w_1, w_2, \ldots, w_k} \frac{1}{2} \| w \|^2 + \frac{1}{N} \sum_{i=1}^{N} \max_{c} \left\{ 1 - w_c^T x_i + w_c^T x_i \right\}
  \]

  \[
  \sum_{i=1}^{N} \sum_{c=1}^{k} \max_{c} \left\{ 1 - w_c^T x_i + w_c^T x_i \right\}
  \]

- Other types of regression/classifications:
  - Extreme value regression (very rare class, treat asychronically)
  - Support Vector regression
  - Ordinal logistic regression, proportional hazards (movie ratings)
  - Ranking (a \succ b)
  - Multiple regression
  - Multi-task classification (predict multiple y's)
  - "Structured prediction"

---

**Convex Functions**

one minimal value (maybe a set of locations)

Why?
- polynomial-time algorithms to optimize
- stationary points are global optima
  \[
  \nabla f(x) = 0
  \]
  (lead sources, L2-roc, L1-roc, logistic, SVMs)

**Convex Set**

A set \( C \) is convex if \( \theta x + (1-\theta) y \in C \)

\[
\forall x, y \in C, \quad 0 \leq \theta \leq 1
\]

i.e., no bending inwards.
Linear Equality  \[ \mathbf{E} \mid a^T \mathbf{x} = b \]  
Inequality  \[ \mathbf{E} \mid a^T \mathbf{x} \leq b \]  
Norm-ball  \[ \{ \mathbf{x} \mid \| \mathbf{x} \| \leq r \} \]  
Norm-cone  \[ \{ \mathbf{x}, r \mid \| \mathbf{x} \| \leq r \} \]  
Spectral cone  \[ \{ \mathbf{E} \mid \mathbf{M} \geq 0 \} \]

\[ \mathbf{E} \mid f(x) \leq 0 \] for any convex \( f \).

Intersection of convex sets is convex

LP  \[ \mathbf{E} \mid A \mathbf{x} \leq b, \mathbf{A} \mathbf{y} = \mathbf{b}, y \geq 0 \]  
"Closed" convex set if it includes boundary

Convex Functions

A function \( f \) is convex if \( \text{dom}(f) \) is a convex set, and
\[ f(\theta x + (1-\theta) y) \leq \theta f(x) + (1-\theta) f(y) \]
\[ \forall x \in \text{dom}(f), y \in \text{dom}(f), 0 \leq \theta \leq 1 \]

Epigraph is a convex set

Examples:
\[ f(x) = ax^2 + b, \quad a > 0 \]  
\[ f(x) = e^x \]  
\[ f(x) = x \log(x), \quad x > 0 \]