Exercises
Advanced Machine Learning
Fall 2019

Machine Learning Laboratory
Dept. of Computer Science, ETH Zürich
Prof. Joachim M. Buhmann
Web https://ml2.inf.ethz.ch/courses/aml/

Series 6. November 27, 2019
(Structural SVMs and Ensemble Methods)

Teaching assistant: **Clara Meister**
meistecl@inf.ethz.ch

Solution 1 (Structural SVMs):

1. It is not immediately obvious why the given $\mathcal{R}(\mathbf{w})$ would be a reasonable objective to drive the learning of the parameters $\mathbf{w}$. When the loss function $\mathcal{L}(z_i, z_j)$ is the 0-1 loss, the objective becomes

$$\mathcal{R}_{\text{0-1}}(\mathbf{w}) = -\log p(\mathbf{w}) + \sum_{i=1}^{n} \log\big(1 - p(z_i \mid \mathbf{y}_i, \mathbf{w})\big).$$

   Contrast this to the usual MAP objective,

$$\mathcal{R}_{\text{MAP}}(\mathbf{w}) = -\log p(\mathbf{w}) - \sum_{i=1}^{n} \log p(z_i \mid \mathbf{y}_i, \mathbf{w}).$$

   While it is clear the maximization of $\log p(z_i \mid \mathbf{y}, \mathbf{w})$ is *related* to the minimization of $\log\big(1-p(z_i \mid \mathbf{y}_i, \mathbf{w})\big)$, one cannot tell that the well-known MAP objective (lecture 4) is recovered for a particular choice of loss function, as it is mistakenly claimed in the textbook that this exercise has been based on [Mur12, Chapter 19.7.1].

   Finally, not even the original authors [YH12, Section 4] of this derivation of the SSVM objective provide more than an intuitive explanation for $\mathcal{R}(\mathbf{w})$ – as minimizing the expected loss, something that we should anyway do according to the Bayes estimator (see paragraph after eq. (7) in the cited paper).

2. This amounts to simple algebraic manipulations, i.e. feeding the definitions of $p(z \mid \mathbf{y}, \mathbf{w})$ and $p(\mathbf{w})$ into $\mathcal{R}(\mathbf{w}) = -\log p(\mathbf{w}) + \sum_{i=1}^{n} \log\left[\sum_z \mathcal{L}(z, z_i) p(z \mid \mathbf{y}_i, \mathbf{w})\right]$, with one notice: the equality required in the exercise sheet holds only up to an additive constant equal to $\log Z$.* The function is generally not convex.

   Notice that the given model is a Conditional Random Field (CRF). See [Mur12, Chapter 19.6]. The $\Psi(z, \mathbf{y})$ function gives the sufficient statistics of the model and $\mathbf{w}$ are the *natural parameters*.

3. We use both inequalities from the hint, once for the function $f_1(z) := \Delta(z, z_i) + \mathbf{w}^T \Psi(z, \mathbf{y}_i)$ and once for

*This is another mistake in the textbook, eq. (19.84).

the function $f_2(z) := \mathbf{w}^T \Psi(z, \mathbf{y}_i)$, with $(\mathbf{y}_i, z_i)$ fixed, to get:

$$\begin{aligned}
\log \sum_z \exp\big(f_1(z)\big) &= \log \sum_z \exp\big(\Delta(z, z_i) + \mathbf{w}^T \Psi(z, \mathbf{y}_i)\big) \\
&\leq \max_z \big\{ \underbrace{\Delta(z, z_i) + \mathbf{w}^T \Psi(z, \mathbf{y}_i)}_{f_1(z)} \big\} + \underbrace{\log |\mathbb{K}|}_{\text{const.}}. \\
\log \sum_z \exp\big(f_2(z)\big) &= \log \underbrace{\sum_z \exp\big(\mathbf{w}^T \Psi(z, \mathbf{y}_i)\big)}_{Z(\mathbf{y}_i, \mathbf{w})} \\
&\geq \max_z \underbrace{\mathbf{w}^T \Psi(z, \mathbf{y}_i)}_{f_2(z)} \\
&\geq \mathbf{w}^T \Psi(z_i, \mathbf{y}_i),
\end{aligned}$$

where the last inequality follows from the definition of $\max$. Hence, the resulting upper bound is

$$\mathcal{R}(\mathbf{w}) \leq E(\mathbf{w}) + \sum_{i=1}^{n} \Big[\max_z \{\Delta(z, z_i) + \mathbf{w}^T \Psi(z, \mathbf{y}_i)\} - \mathbf{w}^T \Psi(\mathbf{y}_i, z_i) \Big] + \text{const.} =: \mathcal{UB}(\mathbf{w}), \tag{1}$$

which is a convex function for certain choices of $E(\mathbf{w})$.

4. For $E(\mathbf{w}) = \frac{1}{2C}\|\mathbf{w}\|^2$ we get the SSVM objective up to a multiplicative factor of C and an additive constant which do not change its minima. The equivalence between the constrained optimization formulation of SSVMs, i.e.,

$$\begin{aligned} &\min_{\mathbf{w},\xi\geq 0} \frac{1}{2}\mathbf{w}^\top\mathbf{w} + C\sum_{i=1}^{n}\xi_i \\ &\text{s.t. } \mathbf{w}^\top\Psi(z_i,\mathbf{y}_i) - \mathbf{w}^\top\Psi(z,\mathbf{y}_i) \geq \Delta(z,z_i) - \xi_i \quad \forall i, \forall z \neq z_i, \end{aligned} \tag{2}$$

and the upper bound from eq. (1) can be seen by fixing $\mathbf{w}$ and solving for ξ_i. This yields the generalized Hinge loss,

$$\xi_i^*(\mathbf{w}) = \max\Big(0, \max_z \Big(\Delta(z,z_i) + \mathbf{w}^T\big(\Psi(z,\mathbf{y}_i) - \Psi(z_i,\mathbf{y}_i)\big)\Big)\Big), \tag{3}$$

where the two $\max$ colapse if we assume that $\Delta(z_i,z_i) = 0$ for all $i = 1,2,\ldots,n$.

5. We have started from an objective function that can be intuitively motivated (but not obtained from some formal assumptions and/or principles), we have defined a conditional random field for our class-posterior distribution and parameter priors, and obtained a convex upper bound by applying two bounding inequalities on the so-called *log-sum-exp* function.

   The beautiful thing about this derivation is that it makes clear the connection between optimizing SSVMs and learning in CRFs. There is a large body of literature on the latter and many of the known methods do not use the same upper bound. This raises interesting questions – which approach is better for a given problem?

Solution 2 (Ensemble Methods):

1. State two essential diferences between bagging and boosting.
   (a) Boosting trains weak learners with the same training set on every iteration, wheareas bagging varies the training sets using resampling.
   (b) Boosting weights the predition of every weak classifier according to its accuracy (during training), whereas bagging gives the same importance to every prediction.
2. How can we use AdaBoost to detect outliers?

   AdaBoost places high weight on samples that are very hard to classify. Therefore, samples with the highest weights might be considered outliers.
3. What is a potential concern with bagging when training on a data set whose labels are highly imbalanced (i.e. one class has many more training points than another). State one measure that can be taken to alleviate this.

   The bootstrapping process could lead to samples where points in the minority class are not present. This can be alleviated by making the size of the ensemble large enough so that at least one point from each class has been sampled in some bootstrap sample.

Solution 3 (Bagging):

A political scientist wants to understand the predictors of household income. She has collected a set of 2000 features on 5000 observations. She wants to use random forests (bagging technique) to model feature interactions but her software implementation is too slow. She comes up with the following plan: she will first run the lasso with cross-validation and extract the predictors P with nonzero coefficients in the selected model. Then she will use just the set P as her features in the random forest.

- Is this a reasonable method? State why or why not.
  No. The lasso will choose features with strong main effects. However, some features may have weak main effects but strong interactions with other features. These would be important for the random forest but will be missed by her screening procedure.

Solution 4 (Boosting):

1. Gradient boosting for a loss function $L(y, f) = (y - f)^2/2$ would complete the following steps:
   (a) Initialize $\hat{f}_0(\mathbf{x}) = \arg\min_h \sum_{i=1}^n (y_i - h(\mathbf{x_i}))^2$
   (b) For m = 1 to M:
      i. Compute the negative gradient
      $$-g_m(\mathbf{x_i}) = y_i - \hat{f}_{m-1}, i = 1 \dots n$$
      ii. Fit a function h_m to the negative gradient by least squares
      $$\begin{aligned}\hat{h}_m &= \arg\min_h \sum_{i=1}^n (-g_m(\mathbf{x_i}) - h(\mathbf{x_i}))^2 \\ &= \arg\min_h \sum_{i=1}^n (y_i - \hat{f}_{m-1} - h(\mathbf{x_i}))^2\end{aligned}$$
      iii. Find β to minimize the loss
      $$\beta_m = \arg\min_\beta \sum_{i=1}^n (y_i - (\hat{f}_{m-1}(\mathbf{x_i}) + \beta \hat{h}_m(\mathbf{x_i})))^2$$
      $$\beta_m = 1$$
      iv. Update $\hat{f}$
      $$\hat{f}_m(\mathbf{x}) = \hat{f}_{m-1} + \hat{h}_m(\mathbf{x})$$
   (c) Output $\hat{f}_M$
2. The algorithm is iteratively fitting the residuals of the previous approximation function $y_i - f_{m-1}$.

References

[Mur12] Kevin P. Murphy, *Machine learning: A probabilistic perspective*, The MIT Press, 2012.

[YH12] Alan Yuille and Xuming He, *Probabilistic models of vision and max-margin methods*, Frontiers of Electrical and Electronic Engineering **7** (2012), no. 1, 94–106.